Modèles et outils pour la conception et la manipulation de systèmes d'aide à la décision

Franck Ravat

Mémoire présenté en vue de l'obtention du diplôme
d'Habilitation à Diriger des Recherches

Spécialité Informatique

Devant l'Université des Sciences Sociales (Toulouse I)
& l'Institut de Recherche en Informatique de Toulouse (IRIT)

par ***Franck RAVAT***

Modèles et outils pour la conception et la manipulation de systèmes d'aide à la décision

Soutenue le13 Décembre 2007 devant la commission d'examen :

M. Bouzeghoub, Professeur à l'Université Versailles Saint-Quentin-en-Yvelines	Rapporteur
C. Cauvet, Professeur à l'Université Paul Cézanne -Aix Marseille III-	Rapporteur
C. Chrisment, Professeur à l'Université Paul Sabatier -Toulouse III-	Examinateur
M. Schneider, Professeur à l'Université Blaise Pascal -Clermont Ferrand II-	Rapporteur
C. Soulé-Dupuy, Professeur à l'Université Toulouse I	Examinateur
G. Zurfluh, Professeur à l'Université Toulouse I	Directeur de recherche

Institut de Recherche en Informatique de Toulouse (UMR 5505), CNRS, Université Paul Sabatier, INP, Université Toulouse I
Site Université Toulouse I, 2 Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 4

REMERCIEMENTS

Mes premiers remerciements s'adressent à Claude Chrisment et Gilles Zurfluh, responsables de l'équipe Systèmes d'Informations Généralisées (SIG), pour m'avoir accueilli au sein de leur équipe et pour m'avoir offert d'excellentes conditions pour mener à bien mes activités de recherche. Qu'ils soient ici remerciés pour leur aide, leur soutien et leurs précieux conseils. J'adresse une mention particulière à Gilles qui a dirigé mes recherches depuis mon DEA. Pendant ces quelques années, j'ai pu apprécier sa disponibilité et sa rigueur dans le travail sans oublier la grande liberté d'action et la confiance qu'il m'a accordées. Merci également pour ses encouragements, son aide de tous les instants et sa sympathie.

Je remercie très sincèrement

- Mme Corine Cauvet, Professeur à l'Université Paul Cézanne -Aix Marseille III-,
- M. Bouzeghoub, Professeur à l'Université Versailles Saint-Quentin-en-Yvelines,
- M. Schneider, Professeur à l'Université Blaise Pascal -Clermont Ferrand II-,

pour avoir accepté d'être rapporteur de ce mémoire, pour leurs remarques pertinentes et pour l'honneur qu'ils me font en participant au jury.

Je tiens également à remercier très sincèrement, Chantal Soulé-Dupuy a plus d'un titre. Tout d'abord, je la remercie pour son aide, son soutien, ses remarques et ses précieux conseils qui ont permis d'améliorer la qualité de ce mémoire. De plus, je tiens à lui exprimer toute ma reconnaissance pour la confiance qu'elle m'a accordée en me donnant différentes responsabilités au sein du master « Ingénierie et Gestion des Systèmes d'Information » puis sa direction depuis quelques années. En tant que directrice de l'UFR Informatique de l'Université Toulouse I, j'apprécie énormément toute l'attention et la considération qu'elle accorde à l'ensemble des responsables de diplômes. Enfin, je la remercie pour l'honneur qu'elle me fait en participant au jury.

Mes remerciements s'adressent également aux "filles" du bureau ME 307Bis, Nathalie Vallès et Geneviève Pujolle, qui par leurs lectures et leurs remarques ont permis d'améliorer la qualité de ce manuscrit. Je n'oublie pas non plus la bonne ambiance et les conversations plus ou moins sérieuses que nous menons avec passion.

Ces remerciements n'auraient pas de sens, si j'oubliais toutes les personnes avec qui j'ai pu collaborer durant ces dix dernières années et notamment au sein de la composante « Entrepôt de Données » de l'équipe SIG. En premier lieu, je tiens à remercier Olivier Teste avec qui c'est un réel plaisir de travailler. Je tiens à le remercier pour tout le travail accompli, son soutien, ses remarques mais également la bonne humeur qu'il sait entretenir au sein de notre bureau de l'IRIT. Je tiens à remercier pour leur précieux travail tous les thésards que j'ai pu co-encadrer (Olivier, Faizza, Kaïs, Estella et Ronan) ainsi que les DEA et/ou Master Recherche. Je n'oublie pas tous les autres membres de l'équipe SIG avec qui j'ai pu collaborer au travers de projets, travaux et/ou articles et que je n'ai pas encore cités (Josiane, Farshad, Max, Guillaume, Gilles, Mohamed…). Je n'oublie pas non plus tous les membres de l'équipe SIG qui rendent les pauses des plus agréables (même si certains jours elles s'éternisent..) voire les repas et notamment les repas de fin d'année. Merci à tous les membres de l'équipe SIG.

Au cours de ces remerciements, je voudrais accorder une mention spéciale à mes collègues de l'Université Toulouse I qui nous permettent de travailler dans des conditions agréables et qui m'ont soutenu et encouragé durant la rédaction de ce mémoire. Je pense notamment aux nombreuses conversations passionnées et passionnantes que nous avons eues durant les pauses déjeuner.

Mes remerciements s'adressent également à tous mes amis qui m'ont soutenu durant ces dernières années. Qu'ils sachent que leur présence est toujours un réel plaisir !

« Last but not least », je voudrais exprimer toute ma gratitude à toute ma famille, et notamment à Maryse pour son soutien de tous les instants et pour ses encouragements. Je tiens à lui dédier ce mémoire ainsi qu'à mes deux garçons qui n'ont pas toujours eu toute l'attention qu'ils sont en droit d'attendre d'un mari ou d'un papa.

SOMMAIRE

SOMMAIRE

INTRODUCTION **1**

CHAPITRE I : CONTEXTE ET PROBEMATIQUE DE NOS TRAVAUX **7**

1 AIDE A LA DECISION **9**
1.1 Système d'Information d'Aide à la Décision (SIAD) 10
1.2 Des systèmes d'aide à la décision aux entrepôts de données 11
1.3 Entrepôt et magasins de données 12
1.4 Bilan sur l'aide à la décision 12

2 ENTREPOTS DE DONNEES FACTUELLES ET DOCUMENTAIRES **13**
2.1 Entrepôts de données factuelles 13
2.2 Entrepôts de documents 14
2.3 Bilan sur les entrepôts 18

3 MAGASINS DE DONNEES OLAP **18**
3.1 OLAP : On-Line Analytical Processing 18
3.2 Modélisation multidimensionnelle 20
3.3 Manipulations de données multidimensionnelles 23
3.4 Méthodes de conception et développement 24
3.5 Bilan sur les magasins de données 26

4 POSITIONNEMENT DE NOS RECHERCHES **26**
4.1 Modélisation des entrepôts 26
4.2 Modélisation des magasins de données multidimensionnelles 27
4.3 Manipulation de données multidimensionnelles 27
4.4 Méthode de conception 28
4.5 Présentation de nos résultats 28

CHAPITRE II : MODELISATION D'ENTREPOTS **29**

1 INTRODUCTION A LA MODELISATION D'ENTREPOTS **31**

2 ENTREPOT DE DONNEES EVOLUTIVES **32**
2.1 Problématique 32
2.2 Concepts 33
2.3 Formalismes et exemples 37

3 ENTREPOTS DE DOCUMENTS **39**
3.1 Problématique 39
3.2 Modèle générique d'entrepôt de documents 40
3.3 Processus d'alimentation 41
3.4 Phase d'extraction 43
3.5 Phase de comparaison 44

4 BILAN ET PERSPECTIVES **47**
4.1 Bilan sur la modélisation des entrepôts 47
4.2 Production scientifique 48
4.3 Perspectives 49

CHAPITRE III : MODELISATION DE MAGASINS DE DONNEES 51

1 INTRODUCTION A LA MODELISATION DES MAGASINS DE DONNEES 53

2 MODELE GENERIQUE DE BASE 54
2.1 Problématique 54
2.2 Dimension 55
2.3 Hiérarchie 55
2.4 Fait 56
2.5 Constellation 57
2.6 Table multidimensionnelle 58

3 INTEGRATION DE DONNEES TEXTUELLES 59
3.1 Problématique 59
3.2 Typologie des mesures 60
3.3 Dimensions représentant un document 60

4 GESTION DE LA COHERENCE SEMANTIQUE 62
4.1 Problématique 62
4.2 Typologie des contraintes 62
4.3 Contraintes sémantiques intra-dimensions 63
4.4 Contraintes sémantiques inter-dimensions 64

5 GESTION DE LA COHERENCE TEMPORELLE 66
5.1 Problématique 66
5.2 Principes 67
5.3 Constellation et versions d'étoile 68
5.4 Composants d'une version d'étoile 69

6 INTEGRATION ET CAPITALISATION DE L'EXPERTISE DES DECIDEURS 70
6.1 Problématique 71
6.2 Principes 71
6.3 Les annotations décisionnelles 72
6.4 Ancrage d'annotations décisonnelles 73

7 PERSONNALISATION DE MAGASINS DE DONNEES 74
7.1 Problématique 75
7.2 Nos résultats 75

8 BILAN ET PERSPECTIVES 77
8.1 Bilan sur la modélisation des magasins 77
8.2 Production scientifique 79
8.3 Perspectives 79

CHAPITRE IV : MANIPULATION DE DONNEES MULTIDIMENSIONNELLES 81

1 INTRODUCTION A L'ANALYSE MULTIDIMENSIONNELLE 83
1.1 Travaux existants 83
1.2 Problématique 83

2 ALGEBRE MULTIDIMENSIONNELLE 85
2.1 Constructeur 85
2.2 Noyau minimum fermé 86
2.3 Opérateurs de second niveau 90
2.4 Opérateurs binaires 91
2.5 Adaptations aux schémas multidimensionnels étendus 93

3 LANGAGE GRAPHIQUE GOLAP 93
3.1 Visualisation d'un schéma multidimensionnel 94
3.2 Création initiale d'une table multidimensionnelle 95
3.3 Manipulations OLAP Graphiques 95
3.4 Complétude du langage graphique GOLAP 96

4 OLAP SQL 96
4.1 Langage de consultation d'OLAP SQL 97
4.2 Langage de définition de OLAP-SQL 98
4.3 Langage de contrôle de OLAP-SQL 99
4.4 Langage de manipulation de OLAP-SQL 100

5 BILAN ET PERSPECTIVES 100
5.1 Bilan sur les langages de manipulation 100
5.2 Production scientifique 102
5.3 Perspectives 102

CHAPITRE V : DEMARCHE DE CONCEPTION 105

1 INTRODUCTION A LA DEMARCHE DE CONCEPTION DE SYSTEME DECISIONNEL 107

2 CONCEPTION DE SCHEMAS MULTIDIMENSIONNELS CONTRAINTS 107
2.1 Problématique 108
2.2 Etapes de la démarche mixte 108
2.3 Approche descendante 109
2.4 Approche ascendante 111
2.5 Confrontation et bilan 114

3 CONCEPTION D'UN SYSTEME D'AIDE A LA DECISION 115
3.1 Problématique 115
3.2 Démarche du Trident Décisionnel 116
3.3 Analyser le SAD 119
3.4 Concevoir le SAD 121
3.5 Principes de réutilisation 124
3.6 Implantation au sein de la société I-D6 126

4 BILAN ET PERSPECTIVES 128
4.1 Conception d'un magasin de données multidimensionnelles 128
4.2 Conception d'un système d'aide à la décision 129
4.3 Production scientifique 129
4.4 Perspectives 130

CHAPITRE VI : OUTILS, PROJETS ET PUBLICATIONS 131

1 INTRODUCTION 133

2 LOGICIEL D'AIDE A LA CONCEPTION GRAPHIQUE D'ENTREPOTS ET DE MAGASIN DE DONNEES 133
2.1 Problématique 133
2.2 Principes 134
2.3 Elaboration graphique et incrémentale d'un entrepôt 135
2.4 Elaboration de magasins multidimensionnels contraints 137

3 OUTIL DE MANIPULATIONS MULTIDIMENSIONNELLES 139
3.1 Architecture 140
3.2 Langage assertionnel 140
3.3 Langage graphique 141

4 CONCEPTION ET MANIPULATION D'ENTREPOTS DE DOCUMENTS 143
4.1 Architecture de DOCWARE 144
4.2 Parseur 144
4.3 Moteur OLAP 145

5 PROJETS ET COLLABORATIONS 145
5.1 Vision synthétique 146
5.2 Présentation des projets et/ou collaborations 146

6 ENCADREMENTS ET PUBLICATIONS 148
6.1 Encadrements 148
6.2 Publications 150

7 BILAN ET SYNTHESE 153

CONCLUSION ET PERSPECTIVES GENERALES 157

BIBLIOGRAPHIE 165

SIGLES 187

INTRODUCTION

1 CONTEXTE

La mondialisation et la concurrence qu'elle engendre rendent le pilotage d'une organisation de plus en plus complexe. Cette complexité est liée non seulement à l'augmentation du nombre de paramètres à prendre en compte mais également à la nécessité de prises de décisions rapides afin d'être réactifs à l'évolution de la concurrence et de la demande des clients. L'efficacité de ces prises de décisions repose sur la mise à disposition d'informations fiables, pertinentes et d'outils facilitant cette tâche. Les systèmes traditionnels, dédiés à la gestion quotidienne d'une organisation, s'avèrent inadaptés à une telle activité [Codd et al., 1993 ; Inmon, 1996 ; Kimball & Ross, 2002]. Face à ce besoin est né le secteur de l'informatique décisionnelle.

Dès le début des années 90, des éditeurs de logiciels ont proposé des outils facilitant la prise de décision. Par exemple, la société Business Objects SA a proposé des logiciels permettant d'effectuer des requêtes graphiques en faisant abstraction des caractéristiques d'implantation des données sources. De nos jours, il existe de nombreux outils permettant d'effectuer des extractions et des transformations de données sources pour les exploiter à des fins décisionnelles. De la même façon, nous retrouvons une multitude d'outils permettant d'effectuer de simples tableaux de bord ou des analyses décisionnelles interactives voire prédictives (Data-mining). Les éditeurs de logiciels tels que Business Objects, Cognos, Microsoft, Oracle, Hypérion sont unanimement reconnus. Ce secteur est toujours en pleine mutation. De nombreuses sociétés, éditrices de logiciels, en rachètent d'autres pour couvrir plusieurs aspects du processus décisionnel. Des sociétés de services se spécialisent dans ce domaine ou créent des départements dédiés au décisionnel. Enfin, les professionnels de l'informatique décisionnelle créent des communautés. A titre d'exemple, Decideo.fr constitue la communauté francophone des professionnels de l'informatique décisionnelle.

Depuis une décennie, même si ce secteur est largement dominé par les outils du marché, l'aide à la décision devient un axe de recherche à part entière[1] [Widom, 1995 ; Chaudhuri & Dayal, 1997] au travers de nouveaux concepts tels que les entrepôts de données – Data Warehouse – [Kimball & Ross, 2002] ou les systèmes On-Line Analytical Processing – OLAP – [Codd et al., 1993]. Durant cette dernière décennie, cette maturité s'est matérialisée par la création de conférences et de journaux dédiés. Nous pouvons notamment citer la conférence internationale DAWAK (International Conference on Data Warehousing and Knowledge Discovery) dont la neuvième édition a eu lieu en septembre dernier, sans oublier la conférence nationale du domaine dont notre équipe est un des membres actifs ; il s'agit d'EDA (journées francophones sur les Entrepôts de Données et l'Analyse en ligne) dont nous organiserons la quatrième édition en Juin prochain. Enfin, nous pouvons citer la revue internationale de référence du domaine "International Journal of Data Warehousing and Mining" (IJDWM[2]) créée en 2005.

De nos jours, un **Entrepôt de Données** (ED), est reconnu comme le composant essentiel des systèmes d'aide à la décision [List et al., 2002 ; Shim et al., 2002] garantissant la meilleure réponse aux problématiques décisionnelles des différents domaines fonctionnels d'une entreprise [Franco & De Lignerolles, 2000]. Bill Inmon définit un ED comme "une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse" [Inmon 1996].

Les premiers travaux sur les ED se sont centrés sur les processus d'intégration, de transformation et de stockage des données sources. Ces premières recherches ont abouti au concept de vues matérialisées [Gupta & Mumick, 1995 ; Widom, 1995]. Une **vue matérialisée** consiste à calculer une vue exprimée sur une source de données et à stocker physiquement les données obtenues dans l'ED. Nous pouvons signaler le projet WHIPS[3] dont l'objectif est la création et la maintenance

[1] Bibliographies disponibles sur Internet : http://www.daniel-lemire.com/OLAP/index.html, http://www-db.stanford.edu/warehousing/publications.html, http://www-lsi.upc.es/~aabello/references.html

[2] http://www.igi-pub.com/journals/details.asp?ID=4291

[3] http://infolab.stanford.edu/warehousing/index.html

efficace d'un ED à base de vues matérialisées. Tous ces travaux sur les vues matérialisées se soucient essentiellement des processus d'alimentation d'un ED (vision physique ou logique) mais ne proposent pas de solution pour la conception de ces nouveaux systèmes d'aide à la décision.

Cependant, l'élaboration d'un ED reste une tâche complexe car elle nécessite l'analyse des besoins de différentes unités organisationnelles d'une entreprise [Bruckner et al., 2001b ; Mazon et al., 2005a] et constitue une nouvelle discipline n'offrant pas des stratégies et des techniques reconnues [List et al., 2002]. De plus, par opposition aux applications de gestion, les ED doivent intégrer la liaison avec les données sources hétérogènes et réparties, proposer une gestion des évolutions de valeurs dans le temps, et supporter une modélisation facilitant les interrogations et les analyses décisionnelles. Sachant que les méthodes de conception dédiées aux applications de production ne sont pas adaptées pour le développement d'applications décisionnelles [Golfarelli & Rizzi, 1998], tous les acteurs sont demandeurs d'**outils méthodologiques** facilitant la conception et le développement de telles applications. A titre d'exemple, la société I-D6, spécialisée dans l'informatique décisionnelle, a fait appel à nos compétences pour définir une démarche de conception d'ED dans le cadre d'un financement CIFRE.

Suite à mes travaux de thèse ayant apporté une solution pour la conception d'applications reposant sur des bases de données réparties [Ravat, 1996 ; Ravat et al., 1997] dès mon recrutement en tant que maître de conférences, j'ai élargi mes travaux vers le thème de la conception de systèmes d'aide à la décision. Plus précisément, les travaux que j'ai menés ces dernières années visent à apporter des outils méthodologiques facilitant la **conception** de tels systèmes ainsi que leur **manipulation**.

2 ORIENTATIONS DE NOS TRAVAUX

Une des premières difficultés de ces travaux est qu'il existe de nombreux outils et concepts sans qu'il y ait de définitions unanimement reconnues par les communautés scientifique et professionnelle [Rizzi et al., 2006]. En effet, suivant leurs préoccupations et leurs fonctions (techniques ou fonctionnelles), les personnes peuvent donner des sémantiques différentes aux composants d'un système décisionnel. Par opposition à certains domaines comme les bases de données relationnelles par exemple, le terme d'entrepôt de données ou "data warehouse" n'a pas toujours la même signification d'un article de recherche à l'autre. Une des premières tâches que nous nous sommes assignés a été de définir les différents concepts servant de support à notre activité de recherche.

D'après la définition de [Inmon, 1996], l'entrepôt de données doit permettre d'extraire, de transformer et de stocker un grand volume de données opérationnelles et, en même temps, de répondre à des requêtes utilisateurs concernant un sujet d'analyse spécifique. En fait, cette définition regroupe deux problématiques différentes. La première permet d'étudier l'intégration des données sources (centralisation, stockage et conservation de l'évolution de l'ensemble des informations sources nécessaires aux prises de décision). L'espace de stockage dédié à ces fonctions est qualifié d'Entrepôt de Données (ED). Ce dernier sert de support à l'élaboration d'espaces dédiés à un métier ou une analyse décisionnelle particulière ; on parle de Magasins de Données (MD) [Ravat et al., 1999]. Cette dichotomie, maintenant reconnue par la communauté du décisionnel, a servi de ligne directrice pour nos recherches.

Dans un premier temps, notre objectif est de proposer comme premiers outils méthodologiques des modèles pour ces différents espaces de stockage des données décisionnelles.

Au niveau de l'ED, il est nécessaire de proposer une solution permettant de stocker l'ensemble des données provenant des sources hétérogènes et réparties. Ce modèle doit intégrer des composants comprenant une partie statique (structure des données décisionnelles) et une partie dynamique traduisant les processus d'extraction et de transformation des données sources. D'autre part, ce modèle d'ED doit permettre de représenter des données calculées à partir de sources mais également leurs différentes évolutions (détaillées ou archivées) au cours du temps. De plus, notre solution doit permettre d'intégrer aussi bien des données factuelles provenant de bases de données de production

que de documents sources. En effet, les ED construits uniquement à partir de BD sources n'exploitent que 20% des informations disponibles alors que les 80% restants sont stockés dans des documents [Tseng & Chou, 2006]. Enfin, une des problématiques majeures dans l'intégration de données non fortement structurées comme des documents est la prise en compte de composants nécessaires aux prises de décision aussi divers que le contenu, les méta-données et la structure. Afin d'intégrer différents documents, notre objectif est de ne pas imposer de structures *a priori*.

La seconde étape de nos travaux consiste à proposer une modélisation adéquate des données d'un MD. Comme indiqué dans [Rizzi et al., 2006] et [Niemi et al., 2003], la modélisation OLAP des données ne repose pas sur une formalisation précise, stable et reconnue par l'ensemble de la communauté scientifique. L'ensemble des propositions reste parcellaire. Aussi, nous souhaitons proposer un modèle générique de données multidimensionnelles afin de pouvoir intégrer l'ensemble des concepts disséminés dans différentes propositions. Ce modèle de base doit être ouvert afin d'intégrer différentes extensions. Ces extensions doivent assurer une modélisation fiable des données décisionnelles et faciliter les prises de décisions.

Afin d'assurer aux décideurs des analyses fiables, le modèle doit intégrer des mécanismes permettant de gérer la cohérence sémantique et temporelle des données. La fiabilité sémantique permet d'assurer une alimentation de données correctes et une manipulation cohérente des données. D'autre part, en complément à notre modèle d'entrepôt, ce modèle doit être capable d'intégrer des indicateurs numériques ainsi que des indicateurs non nécessairement numériques extraits de documents sources. La fiabilité temporelle permet d'intégrer et de sauvegarder les différentes évolutions de valeurs et de structures des schémas multidimensionnels.

Même si les données issues des sources sont fiables, elles peuvent s'avérer insuffisantes pour les prises de décision. Notre objectif est que ce modèle puisse intégrer des informations complémentaires fournies par les décideurs. La sauvegarde du patrimoine immatériel (annotation, commentaire, question, réponse) dans un même espace de stockage faciliterait la tâche du décideur (lecture active) ou d'un ensemble de décideurs devant échanger et confronter leurs avis pour donner une conclusion collégiale. De plus, afin de faciliter les analyses décisionnelles, il doit être possible de personnaliser un schéma particulier afin de préciser les données les plus représentatives pour un décideur.

Afin de compléter ce travail sur la modélisation multidimensionnelle des données, nous souhaitons proposer des solutions pour les manipulations OLAP. A l'instar des modèles, les propositions actuelles s'avèrent parcellaires. Aussi, notre objectif est de proposer une solution complète. Dans un premier temps, nous souhaitons proposer une algèbre orientée utilisateur afin de spécifier l'ensemble des opérations effectuées par les décideurs lors de leurs processus exploratoires des données décisionnelles. Une algèbre n'étant pas destinée à être directement manipulée par les décideurs, nous souhaitons leur proposer à la fois un langage graphique et un langage assertionnel reposant sur les principes formels définis aux travers des primitives algébriques.

Enfin, le dernier aspect que nous souhaitons aborder est relatif aux démarches de conception et de développement de systèmes décisionnels voire d'outils d'aide à la conception. Cette méthode doit permettre d'intégrer aussi bien les besoins des divers décideurs que les sources de données nécessaires à l'élaboration d'un système d'aide à la décision.

Ces différents éléments ont constitué les points fondateurs de mes travaux de recherche de ces dernières années.

3 PLAN DU MEMOIRE

Ce mémoire est composé de 6 chapitres.

Le premier vise à présenter et à définir les différents concepts servant de support à nos travaux. Nous étudions les composants d'un Système d'Information d'Aide à la Décision (SIAD), d'un système d'Aide à la décision (SAD), d'un Entrepôt de données (ED) et d'un Magasin de Données (MD). Nous analysons également les différents travaux associés à chacun de ces composants.

Le second chapitre permet de présenter nos travaux relatifs à la modélisation des entrepôts. Cet espace de stockage de données sources nécessaires aux processus d'aide à la décision doit permettre de stocker les données évoluant dans le temps de manière détaillée ou archivée. Ce modèle d'entrepôt doit également permettre d'extraire et de sauvegarder les informations contenues dans les documents.

Le troisième chapitre présente nos solutions pour une modélisation multidimensionnelle des données d'un MD. Ce modèle permettra de résoudre différents problèmes : définir un modèle générique de base, assurer les fiabilités sémantique et temporelle des données, faciliter les prises de décision au travers de l'intégration de commentaires et de la personnalisation des données.

Le quatrième chapitre permet de présenter les différents langages supportés par ce modèle multidimensionnel. Afin d'être le plus exhaustif possible, nous avons défini un langage procédural à base d'une algèbre multidimensionnelle, un langage déclaratif de type SQL étendu et un langage graphique adapté aux utilisateurs finaux.

Le cinquième chapitre relate notre expérience en matière de démarche d'analyse et de conception des systèmes d'aide à la décision et des magasins de données OLAP.

Le sixième chapitre permet de finaliser la présentation de nos travaux. Notamment, il permet d'expliciter les principales fonctionnalités des outils que nous avons développés, de lister les projets auxquels nous avons participé ainsi que la liste des publications associées à ces travaux.

CHAPITRE I : CONTEXTE ET PROBLEMATIQUE DE NOS TRAVAUX

PLAN DU CHAPITRE

1 L'AIDE A LA DECISION 9
1.1 Système d'Information d'Aide à la Décision (SIAD) 10
1.2 Des systèmes d'aide à la décision aux entrepôts de données 11
1.3 Entrepôt et magasins de données 12
1.4 Bilan sur l'aide à la décision 12

2 ENTREPOTS DE DONNEES FACTUELLES ET DOCUMENTAIRES 13
2.1 Entrepôts de données factuelles 13
2.1.1 Maintenance incrémentale des vues 13
2.1.2 Configuration de l'entrepôt 14
2.2 Entrepôts de documents 14
2.2.1 Vues et Typologie des documents 14
2.2.2 Des documents aux entrepôts de documents 16
2.2.3 Travaux sur les entrepôts de documents 17
2.2.4 Bilan sur les entrepôts de documents 18
2.3 Bilan sur les entrepôts 18

3 MAGASINS DE DONNEES OLAP 18
3.1 OLAP : On-Line Analytical Processing 18
3.2 Modélisation multidimensionnelle 20
3.2.1 Niveau conceptuel 21
3.2.1.1 Extension des modèles existants 21
3.2.1.2 Modèles spécifiques 22
3.2.2 Niveau logique 22
3.2.2.1 R-OLAP : Relational – On Line Analytical Processing 22
3.2.2.2 Autres modèles 23
3.2.3 Niveau physique 23
3.3 Manipulations de données multidimensionnelles 23
3.4 Méthodes de conception et développement 24
3.4.1 Méthodes ascendantes 25
3.4.2 Méthodes descendantes 25
3.4.3 Méthodes mixtes 25
3.5 Bilan sur les magasins de données 26

4 POSITIONNEMENT DE NOS RECHERCHES 26
4.1 Modélisation des entrepôts 26
4.2 Modélisation des magasins de données multidimensionnelles 27
4.3 Manipulation de données multidimensionnelles 27
4.4 Méthode de conception 28
4.5 Présentation de nos résultats 28

Depuis une décennie seulement, l'aide à la décision est devenue un axe de recherche à part entière [Widom, 1995 ; Chaudhuri & Dayal, 1997]. Par opposition à certains domaines comme les bases de données relationnelles, il n'y a pas de définitions unanimement reconnues par les communautés scientifique et industrielle [Rizzi et al., 2006]. Nous retrouvons de nombreux outils sur le marché avec des dénominations très différentes même s'ils assurent des fonctions similaires d'extraction de données ou de restitutions de données décisionnelles. D'autre part, le terme d'entrepôt de données n'a pas toujours la même signification d'un article de recherche à l'autre.

L'objectif de ce chapitre est de définir les concepts ayant servi de support à nos travaux ainsi que notre problématique de recherche. Dans une première section, nous introduisons l'approche d'aide à la décision. Dans les deux sections suivantes, nous étudions les travaux relatifs aux entrepôts de données et aux magasins de données OLAP. Enfin, dans une dernière section, nous précisons les axes de recherche que nous avons suivis durant ces dernières années.

1 AIDE A LA DECISION

La modélisation systémique de toute organisation se décompose en trois sous-systèmes : Système Opérant (SO), Système d'Information (SI) et Système de Pilotage (SP). Le SO représente l'activité productrice de l'organisation étudiée. Cette activité consiste à transformer les flux primaires (matières, finance, personnel…) pour répondre aux besoins des clients. Le SP regroupe l'ensemble du personnel d'encadrement qui effectue les tâches de régulation, de pilotage et d'adaptation de l'organisation à son environnement [Mélèse, 1972]. Le SI permet de collecter, mémoriser, traiter et restituer les différentes données de l'organisation afin de permettre au SP d'effectuer ses fonctions tout en assurant son couplage avec le SO [Nanci & Espinasse, 2001]. L'activité du SO produit des informations stockées dans le SI ; après traitement, la transmission de ces informations vers le SP permet à ce dernier de connaître l'activité du SO (flèches "informations" dans la Figure 1). Les décisions du SP seront répercutées vers le SI puis vers le SO pour permettre au SP d'en maîtriser le fonctionnement (flèches "décisions" dans la Figure 1).

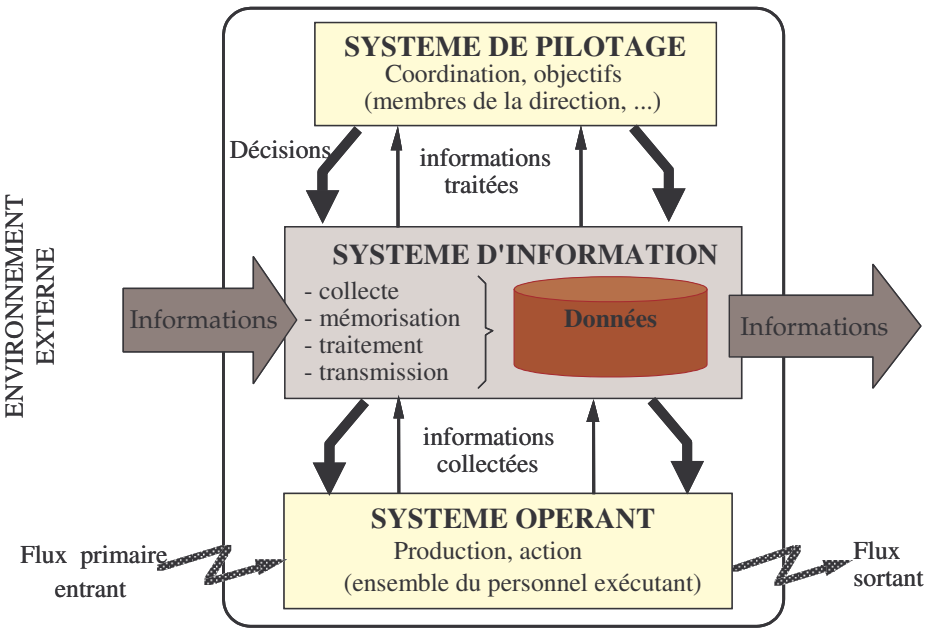


Figure 1: Représentation systémique d'une organisation [Mélèse, 1972]

Pour répondre aux besoins des décideurs, il est nécessaire de synthétiser, réorganiser et historiser les données de production du SI afin d'en déterminer une sous-partie relative à l'aide à la décision. La suite de ce mémoire se centre sur cet aspect. Notamment, dans les sections suivantes, nous définissons les concepts de système d'information d'aide à la décision, de système d'aide à la décision, d'entrepôts et de magasins de données.

1.1 SYSTEME D'INFORMATION D'AIDE A LA DECISION (SIAD)

Par analogie à la définition précédente d'un SI, nous proposons la définition du Système d'Information d'Aide à la Décision (SIAD) suivante :

Définition : Un **SIAD** est la partie d'un système d'information permettant d'accompagner les décideurs dans le processus de prise de décision. Les fonctions d'un SIAD permettent de
- collecter, intégrer, synthétiser et transformer les données opérationnelles d'un SI,
- mémoriser de manière adaptée les données décisionnelles,
- traiter ces données (alimentation, rafraîchissement, pré-calculs…),
- restituer de manière appropriée ces données afin de faciliter la prise de décision.

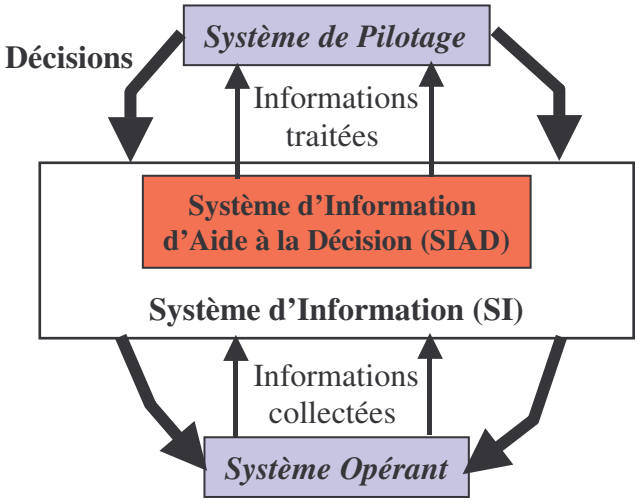


Figure 2 : Le SIAD dans le SI

De nos jours, l'ensemble des outils informatiques permettant de supporter un SIAD est qualifié de Business Intelligence (BI) ou de **Système d'Aide à la Décision (SAD)**. Un SAD vise à exploiter les données opérationnelles d'une organisation afin de faciliter la prise de décision pour un pilotage éclairé. Afin d'être plus explicite, nous proposons la définition suivante :

Définition : Un **Système d'Aide à la Décision** (SAD) regroupe l'ensemble des outils informatiques (matériels et logiciels) permettant :
- d'extraire, de transformer et de charger les données opérationnelles,
- de constituer un ou des espaces de stockage de données décisionnelles,
- de manipuler ces données au travers d'outils d'analyse ou d'interrogation destinés au pilotage des organisations.

La plupart des travaux déclinent ces applications informatiques en trois catégories :

- extraction, transformation et chargement (ou ETL acronyme de "Extraction Transformation Loading") des données opérationnelles (hétérogènes et disparates) pour alimenter et rafraîchir le système d'aide à la décision,
- stockage et traitement des données décisionnelles,
- restitution des données sous une forme adaptée aux utilisateurs (interrogations ou analyses décisionnelles).

Nous pouvons schématiser ces différents outils dans la figure suivante :

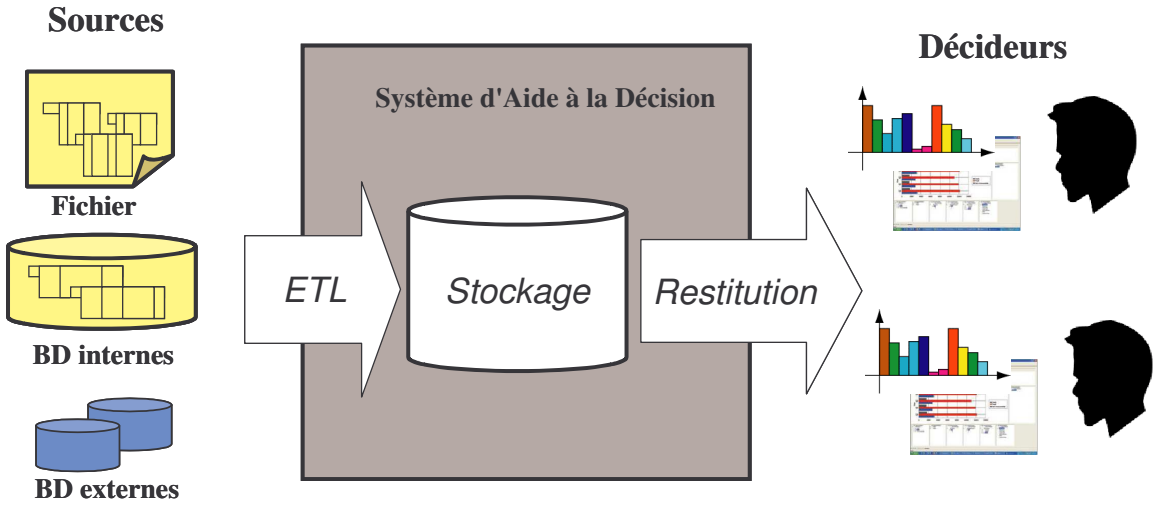


Figure 3 : Le système d'Aide à la Décision

1.2 DES SYSTEMES D'AIDE A LA DECISION AUX ENTREPOTS DE DONNEES

De nos jours, les entrepôts de données constituent une solution adéquate pour construire un système d'aide à décision [Widom, 1995 ; Inmon, 1996]. Un **entrepôt de données** (ED) est défini comme étant "une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse" [Inmon, 1996]. Cette définition met l'accent sur les caractéristiques suivantes :

- **Intégrées :** les données alimentant l'entrepôt proviennent de sources multiples et hétérogènes. Les données des systèmes de production doivent être converties, reformatées et nettoyées de façon à avoir une vision globale dans l'entrepôt.
- **Orientées sujet :** contrairement aux systèmes de production structurant les données par processus fonctionnel, les données d'un ED s'organisent par thèmes d'analyse. L'intérêt de cette organisation est de disposer de l'ensemble des informations utiles sur un thème, le plus souvent transversales aux structures fonctionnelles et organisationnelles d'une entreprise. Cette orientation sujet permet de mettre en avant les indicateurs de performance pour chaque thème d'analyse.
- **Non volatiles :** après intégration, transformation et synthèse des données opérationnelles dans un ED, les seules actions que peuvent effectuer des décideurs sont des interrogations et des analyses décisionnelles (pas de mise à jour).
- **Historisées :** l'alimentation et le rafraîchissement d'un ED consiste en l'intégration des données opérationnelles à différents points d'extraction. Cette intégration de données à des dates différentes permet de conserver "l'historisation" des données qui est vitale pour toute prise de décision.
- **Résumées :** les informations issues des sources doivent être transformées mais surtout agrégées pour faciliter le processus de prises de décision.
- **Disponible pour l'interrogation et l'analyse :** afin d'améliorer les performances d'une organisation, les décideurs doivent pouvoir consulter et analyser les données contenues dans un ED au travers d'outils interactifs.

1.3 ENTREPOT ET MAGASINS DE DONNEES

D'après la définition de [Inmon, 1996], l'ED doit permettre d'extraire, de transformer et de stocker un grand volume de données opérationnelles et, en même temps, de répondre à des requêtes utilisateurs concernant un thème d'analyse spécifique. En fait, cette définition regroupe deux problématiques que nous avons identifiées comme suit dès le début de nos travaux :

- la gestion efficace des données "historisées", "centralisées" (intégration des sources),
- la définition d'un sous-ensemble de données autour d'un thème particulier afin de répondre aux besoins spécifiques de décideurs.

Aussi, l'architecture des systèmes d'aide à la décision que nous proposons est basée sur une dichotomie d'espaces de stockage : l'entrepôt et les magasins de données [Ravat et al., 1999] [Ravat & Teste, 2000c ; Teste, 2000].

Définition : Un **Entrepôt de Données (ED)** est l'espace de stockage centralisé d'un extrait des sources pertinent pour les décideurs. Son organisation doit faciliter la gestion des données et la conservation des évolutions nécessaires pour les prises de décision.

Définition : Un **Magasin de Données (MD)** est un extrait de l'ED adapté à un thème d'analyse particulier et organisé selon un modèle adapté aux outils d'analyse et d'interrogation décisionnelles.

Dans la figure suivante, nous schématisons l'architecture des SAD telle que nous l'avons définie précédemment [Ravat & Teste, 2000a ; Ravat & Teste, 2000b].

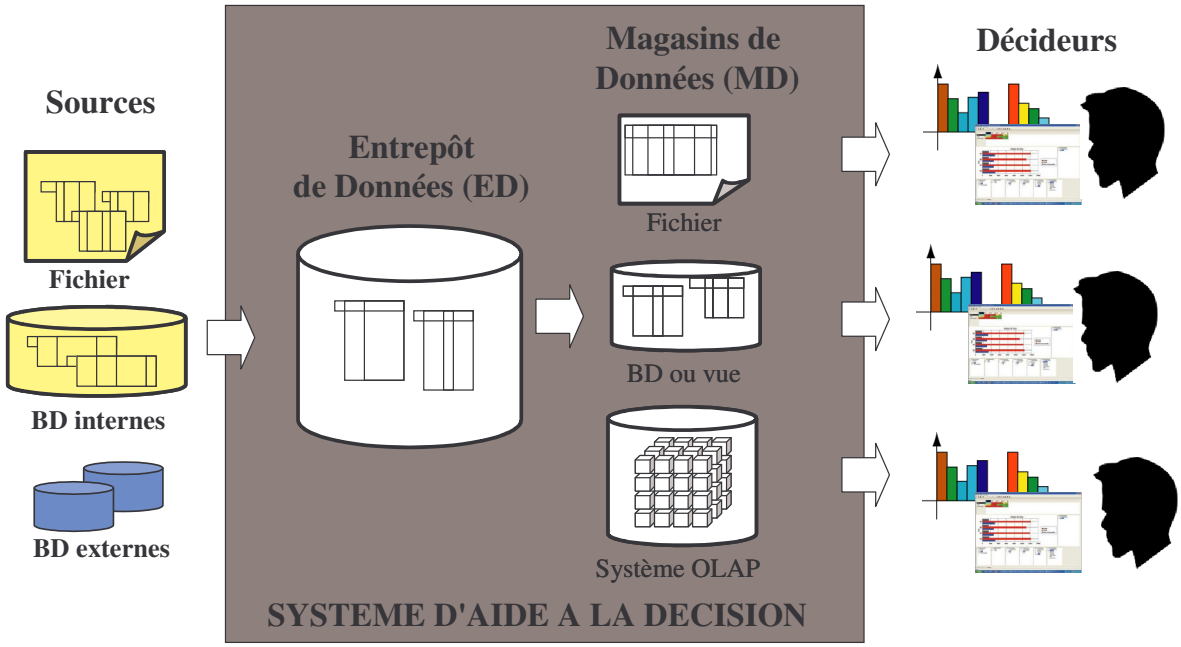


Figure 4 : Entrepôt et magasins de données

Cette dichotomie des espaces de stockage a servi de guide pour nos travaux de recherches.

1.4 BILAN SUR L'AIDE A LA DECISION

Dans cette première section, nous avons défini les différents concepts servant de support à nos travaux sur l'aide à la décision. A partir de la représentation systémique d'une organisation, nous avons identifié le concept de SIAD qui est la partie d'un SI permettant d'accompagner un ou plusieurs décideurs dans le processus de prise de décision. Un Système d'Aide à la Décision

(SAD), partie informatisée d'un SIAD, regroupe l'ensemble des outils informatiques capables d'extraire les données opérationnelles afin de les transformer en informations pertinentes pour les décideurs. D'un point de vue architectural, nous avons identifié deux espaces de stockage des données dans un SAD : l'entrepôt (espace de stockage centralisé) et les magasins de données (espace de stockage extrait d'un ED et centré sur un thème d'analyse particulier).

2 ENTREPOTS DE DONNEES FACTUELLES ET DOCUMENTAIRES

Cette seconde section vise à étudier les différents travaux relatifs aux entrepôts de données. Dès 1995, les premiers travaux se sont centrés sur l'extraction de données factuelles sources afin de pouvoir alimenter et rafraîchir un ED voire directement un MD. Or, les sources étant par nature hétérogènes, des travaux plus récents ont permis d'intégrer au sein des ED des documents. Les deux sous-sections suivantes visent à étudier les travaux relatifs à ces deux problématiques : intégration de données factuelles et intégration du contenu informationnel et structurel d'un document dans un ED.

2.1 ENTREPOTS DE DONNEES FACTUELLES

Les premiers travaux relatifs aux entrepôts de données se sont centrés sur l'**extraction de données factuelles** [Gupta & Mumick, 1995 ; Widom, 1995] à l'aide de la technique des vues matérialisées [Chaudhuri & Dayal, 1997 ; Gupta & Mumick, 1995 ; Widom, 1995]. Une **vue matérialisée** consiste à exprimer une vue au travers d'une requête sur une source de données et à stocker physiquement le résultat de cette requête dans l'entrepôt. Vous trouverez une étude comparative et détaillée de ces travaux dans la thèse d'Olivier Teste [Teste, 2000]. Nous pouvons distinguer deux thèmes de recherche principaux :

- La **maintenance incrémentale** des vues matérialisées qui permet de répercuter immédiatement les mises à jour survenues au niveau des sources de données dans l'ED.
- La **configuration** de l'entrepôt (sélection des vues à matérialiser) qui se propose de déterminer un ensemble de vues à matérialiser dans l'entrepôt de telle sorte que le coût de maintenance soit optimal et ne vienne pas altérer le fonctionnement de l'entrepôt.

Ces travaux abordent les SAD au travers d'un seul espace de stockage sans distinguer l'ED des MD. Dans les sous-sections suivantes, nous utilisons le terme entrepôt comme un terme générique désignant l'espace de stockage du système d'aide à la décision.

2.1.1 Maintenance incrémentale des vues

La majorité des travaux relatifs à la maintenance incrémentale des vues repose sur des entrepôts de données relationnelles [Gupta & Mumick, 1995 ; Quass et al., 1996 ; Hull & Zhou, 1996 ; Zhou et al., 1996 ; Hurtado et al., 1999 ; Huyn, 1996 ; Huyn, 1997 ; Labio et al., 1999 ; Labio & Garcia-Molina, 1996 ; Mumick et al., 1997 ; Quass & Widom, 1997 ; Yang & Widom, 1998 ; Yang & Widom, 2000 ; Zhuge et al., 1995 ; Zhuge et al., 1996 ; Zhuge et al., 1997 ; Zhuge et al., 1998]. Néanmoins, certains travaux reposent sur un modèle objet [Zhuge & Garcia-Molina, 1998]. La maintenance incrémentale est majoritairement effectuée à l'aide de vues matérialisées éventuellement complétées de vues virtuelles [Hull & Zhou, 1996 ; Zhou et al., 1996] ou de vues auxiliaires (vue généralement matérialisée définie pour améliorer le fonctionnement de l'entrepôt [Quass et al., 1996 ; Mumick et al., 1997]). Les vues sont généralement définies au travers d'opérateurs de sélection, de projection et de jointure (SPJ) ou plus rarement d'opérateurs d'agrégation et de regroupement [Hurtado et al., 1999 ; Labio et al., 1999 ; Mumick et al., 1997 ; Quass & Widom, 1997] voire temporels [Yang & Widom, 1998 ; Yang & Widom, 2000]. Les algorithmes de maintenance incrémentale reposent essentiellement sur des modèles de graphes (représentant les plans de décomposition des vues), sur des modèles de coût de maintenance [Labio & Garcia-Molina, 1996] ou de stratégie d'auto-maintenance (maintenance de vues sans

accéder ou en limitant l'accès aux données sources [Quass et al., 1996 ; Huyn, 1996 ; Huyn, 1997 ; Yang & Widom, 1998 ; Yang & Widom, 2000]).

2.1.2 Configuration de l'entrepôt

La configuration d'un entrepôt consiste à déterminer l'ensemble des vues à matérialiser. Ces travaux reposent sur des structures de données relationnelles [Gupta, 1997 ; Gupta & Mumick, 1999 ; Kotidis & Roussopoulos 1999 ; Labio et al., 1997 ; Theodoratos & Sellis 1997 ; Theodoratos & Sellis 1999 ; Yang et al., 1997] ou multidimensionnelles [Baralis et al., 1997 ; Harinarayan et al., 1996 ; Shukla et al., 1998]. Comme pour les travaux précédents, les vues matérialisées sont définies au travers de SPJ complétés éventuellement par des opérateurs d'agrégation ou de regroupement [Baralis et al., 1997 ; Harinarayan et al., 1996 ; Kotidis & Roussopoulos 1999 ; Labio et al., 1997 ; Shukla et al., 1998 ; Yang et al., 1997]. Les algorithmes proposés reposent sur des modèles de graphes et de fonction de coût (intégrant les temps de réponse à l'interrogation, l'espace stockage, le temps de calcul des vues…) voire des stratégies d'auto-maintenance.

D'autres travaux permettent de déterminer les vues à matérialiser en tenant compte de l'évolution du schéma des sources [Bellahsene, 1998], de l'expiration des données (données inutiles dans l'entrepôt ou inadaptées sous leur forme actuelle) [Garcia-Molina et al., 1998], et des relations entre les données de l'ED et les sources dont elles sont issues [Cui & Widom, 2000].

La définition d'un entrepôt peut également reposer sur des facteurs de qualité [Theodoratos & Bouzeghoub, 1999]. Ces facteurs font référence à la complétude des données (toutes les données sources sont présentes pour l'élaboration d'une requête), à la fraîcheur des données (limite temporelle fixée entre la réponse à une requête et la dernière modification de la source), au coût pour la maintenance des vues et des requêtes ainsi qu'à l'intégrité des données. Ces travaux se sont déroulés dans le cadre du projet Esprit DWQ (Data Warehouse Quality) [Jarke & Vassiliou, 1997 ; Vassiliadis et al., 1999].

Tous ces travaux reposant sur le concept de vues matérialisées se soucient essentiellement des processus d'alimentation d'un ED, mais ne proposent pas de solution pour une modélisation adaptée à un large volume de données hétérogènes dont il faut conserver les évolutions.

2.2 ENTREPOTS DE DOCUMENTS

Dans une organisation, de nombreux documents électroniques circulent et peuvent être gérés par des systèmes spécifiques et non nécessairement inter-opérables. Or, ces documents peuvent constituer une source essentielle pour alimenter un entrepôt de données. Aussi, nous avons étudié l'intégration des documents textuels dans un entrepôt de données. Les sous-sections suivantes permettent d'étudier les concepts de document, d'entrepôt de documents et les travaux relatifs aux entrepôts de documents.

2.2.1 Vues et Typologie des documents

Pour alimenter un entrepôt, nous ne retenons que les documents électroniques. Un document électronique est un ensemble d'informations organisées et présentées selon les choix de l'auteur [Fondin, 1998]. Les travaux en recherche d'informations ont permis d'associer plusieurs vues à un même document. Nous pouvons, notamment citer les vues suivantes [Stern, 1997; Sédes, 1998 ; Chrisment et al., 2000 ; Fuhr, 2000] :

- Son **contenu** proprement dit ;
- Sa **structure physique** : cette structure décrit le format (taille, mise en page, formatage) de restitution physique d'un document sur un support (papier, écran…). Cette structure est basée sur la décomposition de son contenu en blocs d'information et une

arborescence de ceux-ci pour la représentation visuelle du document. Un même document peut avoir différentes structures physiques associées à différents supports de visualisation ;

- Sa **structure logique** : cette structure impose d'identifier de manière non ambiguë les granules d'information d'un document [Chrisment et al., 2000] et définit un découpage de l'information d'un point de vue hiérarchique. Une structure logique peut être décomposée en structure générique et structure spécifique [Sèdes, 1998]. La **structure générique** exprime l'organisation générique commune à toute une classe de documents. La **structure spécifique** d'un document est une instance d'une structure générique. Elle est unique et concerne un seul document.
- Ses **attributs externes** ou identité du document [Stern, 1997] : ces attributs permettent de caractériser sans équivoque un document (auteurs, numéro ISBN, année de parution…).

La Figure 5 illustre ces différents composants [Khrouf & Soulé-Dupuy 2004] :

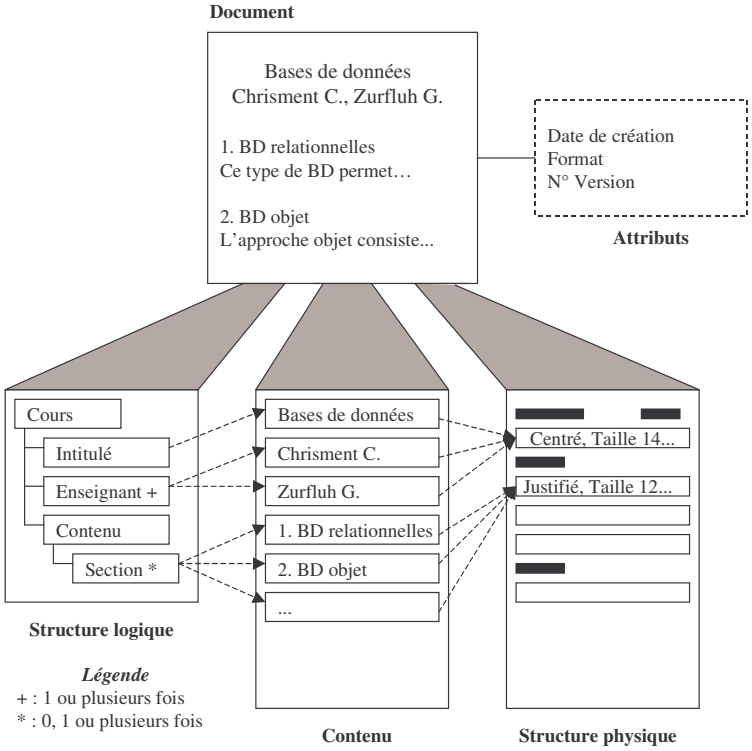


Figure 5 : Vues d'un document

La définition du concept de structure logique a fait apparaître trois types de documents : les **documents non structurés** (contenant peu d'informations de structure), les **documents semi-structurés** (caractérisés par leur structure implicitement déclarée, irrégulière et non rigide) et les **documents structurés** (possédant une structure explicitement déclarée et connue à priori).

Afin de faciliter et d'automatiser l'échange, l'archivage et l'exploitation des documents, des normes reposant sur les structures de documents ont été définies. Notamment, la plus populaire est aujourd'hui XML [W3C, 2000] et ses dérivés définis par le World Wide Consortium (W3C) tels que HTML [W3C, 1999] qui ont aujourd'hui remplacé le langage SGML [ISO, 1986]. De manière globale, ces normes permettent de représenter un document numérique comme un ensemble d'objets plus ou moins complexes et non plus comme une chaîne de caractères.

2.2.2 Des documents aux entrepôts de documents

Selon [Balmisse, 2002], un entrepôt de documents repose sur la centralisation des documents dans un seul espace de stockage. Grâce à cette centralisation, l'extraction des documents sources, la conservation et l'accès aux documents de l'entreprise sont réalisés selon un modèle unique. Cette première définition met essentiellement l'accent sur la centralisation de documents mais néglige l'aspect aide à la décision. De son côté, D. Sullivan [Sullivan, 2001] définit un entrepôt de documents comme un espace de stockage centralisé (1) de documents hétérogènes provenant de sources multiples (2) de caractéristiques essentielles (méta-données, résumé généré éventuellement de manière automatique…) et (3) de relations sémantiques entre documents (index de termes…). Un tel entrepôt sert essentiellement à des analyses orientées « text-mining » [Sullivan, 2001]. Cette définition présente l'inconvénient d'être orientée uniquement vers des analyses textuelles.

Afin de répondre aux limites des propositions précédentes, nous introduisons la définition suivante :

Définition : Un **Entrepôt de Documents (EDO)** est un espace de stockage centralisé d'informations (contenu, structures, métadonnées…) issues des documents sources (hétérogènes en structures et en types) pertinents pour les décideurs. Cet entrepôt sert de support aux processus de recherche d'information, d'interrogation et d'analyse décisionnelles.

Suite à cette définition, la question reste entière quant au stockage des documents dans un EDO. Les EDO devant supporter un volume important d'informations, il est nécessaire de définir des modèles et des outils permettant de stocker les structures et le contenu des documents voire d'organiser le fond documentaire entier afin de faciliter les restitutions décisionnelles.

Une première solution consiste à utiliser un Système de Recherche d'Information (SRI). Un SRI est un ensemble de modèles et de processus permettant la sélection d'informations pertinentes en réponse aux besoins utilisateurs exprimés à l'aide d'une requête contenant le plus souvent des mots clés [Salton, 1971 ; Rijsbergen 1979]. L'architecture d'un SRI comprend donc un module **d'indexation** des documents ou de ses unités informationnelles, un module de **formulation** de requête utilisateur et un module de **comparaison** requête-documents afin de restituer à l'utilisateur "uniquement" des informations pertinentes et "toutes" les informations pertinentes. La qualité de ces systèmes tient en grande partie à son processus d'indexation et à la pertinence du modèle de représentation sous-jacent. Tous ces aspects ont été étudiés dans de nombreux ouvrages [Salton, 1971 ; Rijsbergen 1979 ; Salton, 1989 ; Frakes & Yates, 1992 ; Baeza-Yates & Ribeiro-Neto, 1999 ; Chowdhury, 2004] et de nombreux travaux de recherche et notamment dans l'équipe SIG de l'IRIT [Tuffery, 1984 ; Anton, 1987 ; Denjean, 1989 ; Soulé-Dupuy, 1990 ; Aboud, 1990 ; Boughanem, 2000 ; Mothe, 2000 ; Soulé-Dupuy, 2001].

La problématique essentielle dans l'élaboration d'un EDO est la prise en compte de l'hétérogénéité structurelle des documents textuels sources mais également de l'hétérogénéité sémantique (contenu). L'hétérogénéité structurelle des sources se matérialise par des documents plus ou moins structurés et reposant ou non sur des standards. L'hétérogénéité sémantique est liée au fait que les documents pouvant être intégrés dans un EDO peuvent concerner des domaines très divers. Or, les SRI traditionnels, basés sur des index de termes, n'apportent pas de solutions satisfaisantes. Premièrement, la plupart des SRI utilisent comme moyen de stockage, des systèmes propriétaires, le plus souvent sous forme de fichiers classiques. Cette particularité implique une certaine lourdeur quant à l'intégration de documents hétérogènes dans un EDO comme nous souhaitons le faire ; il faut en effet développer des programmes spécifiques pour chaque type de documents. Deuxièmement, les SRI traditionnels basés sur des index de termes ne permettent pas d'exploiter la structure logique des documents de façon appropriée dans la mesure où le document est considéré comme un ensemble de termes (pondérés ou non) et une collection comme un ensemble de documents non structurés [Soulé-Dupuy, 2001]. La prise en

compte de la structure logique offre l'opportunité d'interroger et restituer des parties (ou granules) d'un document textuel. De plus, l'identification de ces granules peut être effectuée à l'aide de normes (telles que SGML [ISO, 1986], HTML [W3C, 1999], XML [W3C, 2000]). Les SGBD sont alors mieux adaptés à une gestion de ce type d'informations que les SRI [Soulé-Dupuy, 2001]. De plus, la gestion de documents avec un SGBD offre de nouvelles perspectives quant à l'exploitation décisionnelle du contenu et des structures des documents. Les interrogations décisionnelles pourront reposer sur l'utilisation d'un langage ensembliste pour restituer des données factuelles liées aux documents (attributs ou méta-informations).

2.2.3 Travaux sur les entrepôts de documents

Dans le domaine des entrepôts de documents, nous avons recensé quatre travaux majeurs : WIND [Faulstich et al., 1997, 1998], Xylème[4] [Abiteboul et al., 2001, 2002], Whoweda[5] (**W**are**ho**use **o**f **We**b **Da**ta) [Bhowmick et al, 2000a, 2000b] et ODS [Boussaid et al., 2006, 2007].

WIND [Faulstich et al., 1997, 1998] est un outil permettant de construire un entrepôt d'informations issues du Web. Cette approche se limite à un domaine particulier ; l'exemple proposé est la base de données cinématographique du Web (Internet Movie Database). WIND a été conçu pour pouvoir traiter aussi bien les documents structurés que des documents en format texte brut. D'un point de vue physique, les données peuvent être stockées aussi bien dans un SGBD, un système de fichier ou un SRI et le système fournit une vision unifiée au travers d'un schéma reposant sur une structure hiérarchique des données. Même si ce système accepte plusieurs types de documents, il faut définir au préalable la grammaire de chacun de ces types.

Xylème [Abiteboul et al., 2002] est un système de gestion d'entrepôts de données XML. Ce système repose sur un SGBD natif permettant de stocker, classer et indexer des documents XML [Abiteboul et al., 2001]. Xylème repose sur des vues qui intègrent les données du Web par domaines sémantiques et les utilisateurs peuvent interroger les données au travers de ces vues. Ce système est désormais commercialisé. Les utilisateurs peuvent effectuer des requêtes à l'aide de mots clés et de critères de sélection portant sur le contenu et la structure des documents.

Whoweda est un entrepôt construit à partir d'informations issues du Web. Il repose sur une Base de Données (BD) contenant des méta-données et la structure des hyperliens entre les pages Web [Bhowmick et al, 2000a]. Cette BD est accessible via deux modules : Web Information Coupling System (WICS) et Web Information Mining System (WIMS). Notamment, le module WICS extrait les informations du Web, les stocke et offre des mécanismes de manipulation de données via des opérateurs tels que Web Select, Web Join sur le graph représentant les liens entre les pages Web [Bhowmick et al, 2000b].

ODS (Operational Data Store) est un système d'entreposage de données complexes exploitant XML comme langage pivot [Boussaid et al., 2007a]. Ce système repose sur un méta-modèle UML décomposant les objets complexes en différents éléments de base en fonction de leur type (texte, image, son vidéo et vue relationnelle pouvant être matérialisée). Ce méta-modèle contient une classe "Specific" permettant d'instancier les éléments de base et leurs relations sémantiques. Ce méta-modèle peut être traduit ensuite en XML puis en un schéma physique relationnel ou XML. Ce système ODS peut servir de support à l'approche X-Warehousing. Cette approche traduit les besoins utilisateurs en un schéma multidimensionnel XML et fusionne ce dernier avec un schéma de l'ODS exprimé en XML [Boussaid et al., 2007b]. Le schéma résultat sert de support à des analyses basées sur la fouille de données.

[4] www.xyleme.com
[5] http://mandolin.cais.ntu.edu.sg/~whoweda/index.htm

2.2.4 Bilan sur les entrepôts de documents

Les travaux décrits dans cette section sont liés à la prise en compte de la structure logique des documents. Notamment, le modèle de WIND est un modèle adapté à un type de document possédant une structure fortement structurée. De manière générale, leur objectif consiste à décomposer les documents en entités informationnelles afin de les interroger et de les manipuler.

La faiblesse de ces travaux est mise en évidence lorsque la structure logique est absente ou n'est pas spécifiée (textes bruts, textes numérisés), ou lorsque cette structure logique, par son manque de contraintes, permet d'aboutir à des documents trop hétérogènes au niveau de leurs structures (cas des documents Web). Dès lors que la structure n'est pas définie au préalable ou que les règles de structuration ne sont pas rigoureuses, il est difficile d'extraire la structure logique et de faire des interrogations pertinentes pour le pilotage d'une organisation. A l'exception des travaux de [Boussaid et al., 2007a, 2007b], ces différentes propositions négligent l'aspect restitution décisionnelle. Cependant, les travaux de [Boussaid et al., 2007a, 2007b] ne proposent qu'une décomposition d'éléments basiques sans proposer une hiérarchisation des modules d'information. Par exemple, un texte est uniquement caractérisé par un nombre de caractères, un nombre de lignes et un contenu.

2.3 BILAN SUR LES ENTREPOTS

Notre objectif est de proposer des modèles pour des entrepôts comprenant aussi bien des données factuelles que des documents. Dans les travaux existant sur les entrepôts de données factuelles, les travaux se sont essentiellement concentrés sur la définition de vues matérialisées et la maintenance incrémentale de ces dernières. Ces techniques se centrent essentiellement sur les transferts de données entre les sources et un entrepôt cible. La majorité des travaux propose un entrepôt contenant des tables relationnelles. A notre connaissance, il n'y a pas de travaux apportant une solution pour une modélisation conceptuelle d'un entrepôt gérant un grand volume de données dont les décideurs souhaitent conserver l'évolution.

Pour les entrepôts de documents, les solutions proposées reposent sur un modèle de données commun et une ou plusieurs unités de stockage physique. Le plus souvent, les systèmes proposés reposent sur une centralisation contrainte par des schémas définis au préalable. De plus, ils ne proposent pas de solutions pour une restitution à but décisionnel des données contenues dans l'entrepôt.

3 MAGASINS DE DONNEES OLAP

Dans le cadre de nos propositions, nous avons défini deux espaces de stockage pour un système d'aide à la décision : l'entrepôt et les Magasins de Données (MD). Dans cette section, nous nous focalisons sur ce second espace.

Un magasin de données est un extrait de l'entrepôt orienté vers un thème d'analyse particulier et destiné à un type de décideurs. De ce fait, la structuration des données dépend de l'outil utilisé par le décideur lors de ses analyses ou de ses interrogations décisionnelles. De nos jours, les systèmes OLAP sont considérés comme les plus adaptés pour faciliter les prises de décisions [Kimball & Ross, 2002]. Dans cette section, nous définissons le terme OLAP et nous étudions les travaux relatifs à la modélisation multidimensionnelle et aux langages de manipulation OLAP.

3.1 OLAP : ON-LINE ANALYTICAL PROCESSING

Suivant [Pendse, 2006], le terme OLAP, acronyme de On Line Analytical Processing, a été défini en 1993 par EF Codd [Codd et al., 1993], même si certains concepts, technologies voire certains outils ont été élaborés avant. Par opposition à ses propositions du modèle relationnel

reposant sur des principes mathématiques, le rapport technique de EF Codd [Codd et al., 1993], commandité par la société Arbor Software (intitulée actuellement Hyperion Solutions) repose essentiellement sur des considérations commerciales [Pendse, 2005]. Ces règles sont centrées sur les fonctionnalités d'un outil OLAP et non sur les spécificités liées à la modélisation multidimensionnelle des données inhérente à un système OLAP.

Dans ce rapport technique un système est qualifié de OLAP s'il respecte les douze règles énoncées ci-après. Pour en faciliter la compréhension, nous les avons classées en trois catégories :

- Architecture du système OLAP : (1) accessibilité à de nombreuses sources de données, (2) support multi-utilisateurs, (3) architecture client/serveur (4) transparence du serveur OLAP à différents types de logiciel
- Données du système OLAP : (5) vue conceptuelle multidimensionnelle, (6) dimensions génériques (principe de hiérarchisation), (7) nombre illimité de dimensions et de niveaux d'agrégation, (8) gestion dynamique des matrices creuses
- Manipulation de données des systèmes OLAP : (9) manipulation intuitive des données (10) uniformité des performances du système de reporting (restitution), (11) souplesse et facilité de construction des rapports, (12) calcul au travers des dimensions.

En réponse à cette première proposition, les auteurs du "OLAP Report"[Pendse, 2005] ont élaboré le test FASMI pour analyser et classifier les différents outils OLAP :

- **Fast** : le système doit être capable de fournir en quelques secondes les réponses aux requêtes décisionnelles simples comme complexes
- **Analysis** : le système doit être suffisamment flexible pour supporter par défaut différents modèles d'analyse et/ou de calcul statistique et offrir la possibilité aux décideurs de faire des calculs ad hoc.
- **Shared** : le système met en application toutes les conditions de sécurité pour la confidentialité et les mises à jour concurrentes.
- **Multidimensional** : cette fonctionnalité essentielle offre aux décideurs une vision conceptuelle multidimensionnelle avec une hiérarchisation des dimensions d'analyse.
- **Information** : les auteurs étudient les différents types de sources possibles pour alimenter le système OLAP

Les définitions précédentes visent principalement à classifier les outils OLAP du marché. Or, dans le contexte de nos travaux, nous souhaitons faire abstraction des outils du marché et proposer une définition mettant en avant la caractéristique de base des systèmes OLAP : la structure multidimensionnelle des données. Autrement dit, les données forment des points dans un espace à plusieurs dimensions afin de s'approcher de la perception du décideur. Ces points représentent des centres d'intérêts décisionnels (également appelés sujets d'analyse) analysés selon différents axes d'analyse (ou dimensions). Cette représentation multidimensionnelle des données sert de support aux processus d'aide à la prise de décision. Nous proposons alors de définir les systèmes OLAP comme suit :

Définition : Un système **OLAP** permet de stocker et de présenter de manière multidimensionnelle les données décisionnelles. Cette structuration multidimensionnelle des données (indicateurs de performance analysés en fonction de différents axes) sert de support à l'interrogation, à la synthèse dynamique (tableaux de bord) et à l'analyse interactive (ou OLAP).

3.2 MODELISATION MULTIDIMENSIONNELLE

Pour le modèle multidimensionnel, des concepts et des systèmes existent sans fondement théorique stable [Marcel, 1998 ; Niemi et al., 2003 ; Rizzi et al., 2006]. En l'absence d'un modèle consensuel, plusieurs propositions ont été présentées. La plupart de ces modèles reposent sur un sujet d'analyse (fait) associé à des axes d'analyse (dimensions) [Kimball & Ross, 2002].

Afin de pouvoir comparer facilement les travaux existants, nous proposons une définition succincte des concepts du modèle multidimensionnel [Ghozzi, 2004]. Un **fait** représente un centre d'intérêt décisionnel ou autrement dit un sujet d'analyse. Il regroupe un ensemble d'attributs le plus souvent numériques et cumulables appelés **mesures** d'activité ou indicateurs. Les mesures d'un même fait peuvent être analysées suivant différents axes. Un axe d'analyse, également appelé **dimension**, regroupe un ensemble d'attributs et de hiérarchies. Une **hiérarchie** est une perspective ou vision d'analyse définie au sein d'une dimension. Elle regroupe un ensemble d'attributs organisés du niveau de granularité le plus fin vers le niveau de granularité le plus général [Lehner, 1998]. L'attribut qui permet d'identifier un niveau d'agrégation est appelé **paramètre** tandis que les autres attributs de ce même niveau qui complètent la sémantique de ce paramètre sont appelés **attributs faibles**. Un schéma composé d'un fait et de ses dimensions est qualifié de **schéma en étoile** [Kimball & Ross, 2002]. Un **schéma en constellation** permet de généraliser le concept d'étoile en permettant de définir plusieurs faits comprenant des dimensions spécifiques ou partagées [Moody & Kortink, 2000 ; Ravat et al., 2002]. Un **cube de données** est une structure combinant une partie d'un schéma en étoile et les valeurs associées.

Dans l'exemple ci-dessous, nous représentons dans la partie haute, un schéma en étoile composé d'un fait (VENTES) et de 3 dimensions (TEMPS, PRODUITS et MAGASINS). Ce schéma est représenté à l'aide du formalisme proposé par M. Golfarelli [Golfarelli et al., 1998]. Dans la partie basse, un cube de données construit à l'aide de ce schéma permet d'analyser les ventes annuelles des classes de produit en fonction des villes des magasins.

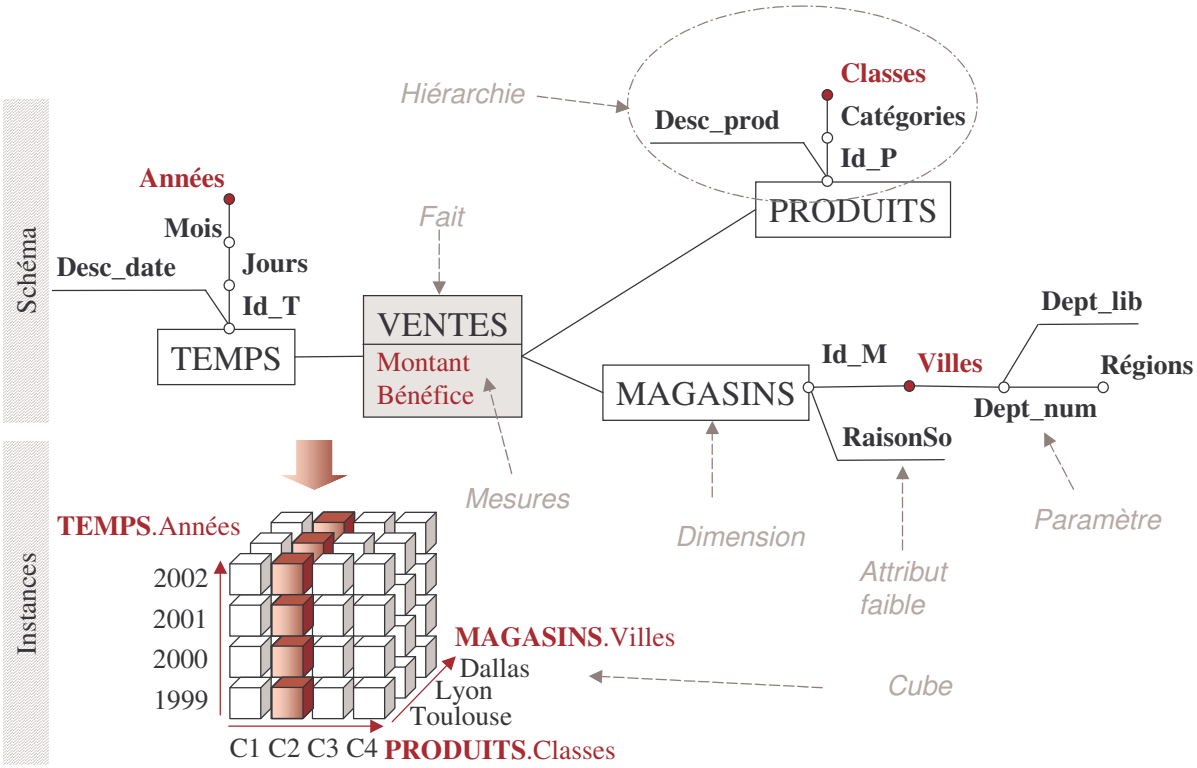


Figure 6 : Schéma en étoile et cube associé

Les états de l'art très complets de [Pedersen et al., 2001], [Torlone, 2003] et [Abelló et al., 2006] permettent de classer les travaux liés à la modélisation multidimensionnelle en deux

catégories. Historiquement, la première catégorie, intitulée "Modèle Cube" [Vassiliadis & Sellis, 1999] permet de représenter les données d'un sujet d'analyse en tant que cellules d'un cube dont les arrêtes regroupent les valeurs des paramètres des dimensions [Agrawal et al., 1995 ; Li & Wang, 1996 ; Gyssens et al., 1996 ; Agrawal et al., 1997 ; Gyssens & Lakshmanan, 1997 ; Datta & Thomas, 1999]. Ces travaux présentent les inconvénients suivants :

- pas de séparation entre la structure et le contenu,
- faiblesse de la modélisation du fait et de ses indicateurs d'analyse,
- intégration d'un seul sujet d'analyse,
- peu ou pas de modélisation de la hiérarchisation des dimensions.

Les travaux de [Lehner, 1998] répondent partiellement à ces limites en proposant une hiérarchisation des données d'une dimension de cube. Néanmoins, afin de pallier ces inconvénients, est apparue la seconde catégorie intitulée "modèle multidimensionnel". Ce modèle, sémantiquement plus riche, permet de définir précisément les différents composants d'un schéma multidimensionnel tels que les faits et les dimensions [Pedersen et al., 2001 ; Abelló et al., 2003 ; Abelló et al., 2006]. Dans ces modèles, apparaissent clairement les différents attributs d'une dimension, classés en hiérarchies représentant les niveaux d'agrégation des indicateurs d'un fait. **De par ces avantages, nous situons nos travaux dans cette seconde catégorie.**

Pour étudier les différentes propositions de la littérature scientifique et même si cet aspect n'est pas toujours mis en avant par leurs auteurs, nous proposons de classifier ces travaux selon les 3 niveaux d'abstractions reconnus et classiquement utilisés (conceptuel, logique et physique).

3.2.1 Niveau conceptuel

Pour le niveau conceptuel, nous proposons d'organiser les modèles multidimensionnels en deux catégories selon le paradigme utilisé : les modèles qui étendent un modèle existant (Entité Association ou orienté objet), et les modèles spécifiquement multidimensionnels.

3.2.1.1 Extension des modèles existants

Afin de faciliter l'acceptation du modèle multidimensionnel, plusieurs auteurs ont choisi de baser leurs travaux sur une approche existante, notamment les modèles Entité-Association ou orienté objet. Ce choix vise à profiter de la popularité de ces modèles ainsi que de la maturité et de la robustesse de leurs concepts et formalismes.

Pour le paradigme Entité-Association [Sapia et al., 1998 ; Tryfona et al., 1999] et [Hahn et al., 2000], une dimension est généralement représentée par un ensemble d'entités. Chaque entité correspond à un niveau hiérarchique et les différents niveaux sont reliés par des associations binaires. Les faits correspondent à des associations entre les différentes dimensions.

Le modèle GOLD proposé par [Trujillo et al., 2003 ; Luján-Mora et al., 2004 ; Luján-Mora et al., 2006] est une extension du diagramme de classes UML. Le fait est représenté par une classe reliée aux classes de dimensions par des liens d'agrégation. Une dimension est représentée par une classe et un graphe acyclique représentant les hiérarchies. Quant aux auteurs de [Abelló et al., 2002], ils proposent un modèle basé sur la spécialisation du méta-modèle d'UML. Dans cette solution (YAM²), les hiérarchies s'expriment par une extension de la relation de composition entre classes [Abelló et al., 2006]. Pour terminer, il faut noter la présence du standard CWM[6] (acronyme de "Common Warehouse MetaModel") qui étend les méta-classes d'UML non pas pour effectuer une modélisation multidimensionnelle mais pour faciliter l'échange de méta-données entre les outils décisionnels et les sources hétérogènes et distribuées [Abelló et al., 2006].

[6] http://www.cwmforum.org/

Ces solutions présentent l'inconvénient de détourner les concepts ayant une sémantique spécifique à des fins de modélisation multidimensionnelle. De plus, ces solutions nécessitent pour un concepteur décisionnel la compréhension d'un besoin décisionnel cumulé à la compréhension de concepts initialement définis pour la modélisation d'applications OLTP (On Line Analytical Processing). Via l'utilisation de stéréotypes, ces propositions conservent la dualité fait-dimension et à l'exception de [Luján-Mora et al., 2006], ces propositions ne profitent pas de la partie dynamique des classes pour notamment modéliser les processus ETL.

3.2.1.2 Modèles spécifiques

Les modèles purement multidimensionnels se basent sur les concepts de fait et de dimension. La simplicité des notations et des concepts ainsi que le souci de se rapprocher de la vision conceptuelle du décideur est le point fort de ces modèles.

Dans cette section, nous pouvons citer les travaux suivants : Dimensionnal Fact Model - DFM – [Golfarelli et al., 1998], EMDM [Pedersen & Jensen, 1999 ; Pedersen et al., 2001], Object Oriented Multidimensionnal Model – OOMDM – [Nguyen et al., 2000 ; Bruckner et al., 2001a], Temporal OLAP – TOLAP – [Hurtado et al, 1999 ; Mendelzon et al, 2003] et le Modèle Dimensionnel (MD) [Cabibbo & Torlone, 1998 ; Cabibbo & Torlone, 2000 ; Torlone, 2003]. Enfin, le modèle OMM-DWG (Object Multidimensional Model - Data Warehouse Graph) permet de représenter de manière uniforme les composants de faits et de dimensions [Schneider, 2003, 2007]. Notamment ces travaux permettent d'effectuer des liens inter-faits et les dimensions peuvent contenir plusieurs entrées pour une connexion à un fait.

Comme tout modèle multidimensionnel, ces travaux proposent une séparation du contenu et de la structure. Cette dernière supporte la définition d'un ou plusieurs faits (TOLAP, MD, OMM-DWG). Les dimensions peuvent contenir une ou plusieurs hiérarchies explicitement représentées sans forcément intégrer le concept d'attribut faible (EMDM, OOMDM, OMM-DWG et TOLAP). Un seul modèle intègre l'expression de contraintes au niveau du modèle multidimensionnel [Hurtado & Mendelzon, 2002]. Certains de ces travaux supportent une représentation graphique des données (DF, OMM-DWG et MD). Notamment, le formalisme graphique de DF étant le plus explicite, nous en avons repris certains composants dans nos propositions.

3.2.2 Niveau logique

Pour une modélisation de niveau logique, il est possible d'utiliser les modèles R-OLAP, O-OLAP, M-OLAP ou H-OLAP.

3.2.2.1 R-OLAP : Relational – On Line Analytical Processing

Le modèle R-OLAP, initialement proposé par [Kimball & Ross, 2002], repose sur les concepts de table de dimension et de table de fait.

Une **table de dimension** est une table dont les attributs correspondent aux paramètres et aux attributs faibles de la dimension ; la clé primaire de cette table correspond au paramètre de granularité la plus faible. Une **table de fait** est une table ayant comme attributs les mesures du fait et les clés étrangères référençant les clés primaires des tables de dimension reliées. La clé primaire d'une table de fait est composée de la concaténation des différentes clés étrangères.

Il est possible de normaliser les tables de dimension pour obtenir le ***schéma en flocons.*** Chaque dimension est transformée en plusieurs tables relationnelles en respectant la troisième forme normale. Ainsi, chaque niveau hiérarchique est transformé en une table ayant comme clé primaire, le transformé du paramètre de ce niveau et comme attributs, les attributs faibles qui lui sont attachés. Chaque table de niveau hiérarchique comporte une clé étrangère référençant la table du niveau suivant dans la hiérarchie.

3.2.2.2 Autres modèles

O – OLAP (Object – On Line Analytical Processing) : ce modèle représente les différents concepts multidimensionnels au travers de classes d'objets. Par exemple, [Buzydlowski et al., 1998] propose un modèle comportant trois catégories d'objets : données, contrôle et interface. Les objets de données sont les faits et les dimensions. Les objets de contrôle sont les requêtes, les opérations OLAP. Enfin, les objets d'interface sont les outils permettant de visualiser les résultats des objets de contrôle.

M – OLAP (Multidimensional – On Line Analytical Processing) : cette approche se base sur un modèle structurant les données dans des cubes de données, des matrices ou des vecteurs à n dimensions [Agrawal et al., 1997 ; Vassiliadis, 1998]. Ces structures optimisent les temps d'accès aux données et réduisent les temps de réponse aux requêtes [Kimball & Ross, 2002].

H – OLAP (Hybrid – On Line Analytical Processing) : ce modèle propose de cumuler les avantages des deux modèles M-OLAP et R-OLAP[7]. Les données agrégées sont stockées sous forme multidimensionnelle, alors que les données détaillées sont stockées dans des tables relationnelles.

3.2.3 Niveau physique

Au niveau physique, plusieurs travaux se sont focalisés sur les techniques de matérialisation des vues (cf. sections 2.1.1 et 2.1.2.), d'indexation (optimisation des temps d'accès aux données) et de fragmentation. Pour la gestion des données OLAP volumineuses et l'optimisation des requêtes, nous pouvons citer les travaux liés aux index binaires [Wu et al., 2004 ; Lim & Kim, 2004], aux index de projection [Gupta et al., 1997] voire à la combinaison des trois techniques (matérialisation de vues, index et fragmentation) [Bellatreche et al., 2004].

3.3 MANIPULATIONS DE DONNEES MULTIDIMENSIONNELLES

Comme indiqué dans les sections précédentes, les modèles multidimensionnels reposent sur la métaphore du cube de données matérialisé au travers des concepts de fait et de dimension. Les premiers travaux ont adapté les opérateurs de l'algèbre relationnelle aux cubes de données [Gray et al., 1996 ; Agrawal et al., 1997 ; Li & Wang, 1996 ; Gyssens & Lakshmanan, 1997 ; Datta & Thomas, 1999 ; Rafanelli, 2003]. Notamment, les auteurs de [Gray et al., 1996] ont proposé une extension du langage SQL avec l'opérateur « DataCube » qui généralise le "group by" pour supporter des analyses multidimensionnelles.

Par opposition aux interrogations transactionnelles, les manipulations de données multidimensionnelles consistent à appliquer des processus exploratoires afin de spécifier le cube ou le tableau résultat à *N* dimensions [Tininini, 2003]. Pour effectuer ces processus exploratoires, il est proposé différentes opérations [Li & Wang, 1996 ; Agrawal et al., 1997 ; Cabibbo & Torlone, 1997 ; Gyssens & Lakshmanan, 1997 ; Cabibbo & Torlone, 1998 ; Lehner, 1998 ; Marcel, 1998 ; Pedersen & Jensen, 1999 ; Abelló et al., 2003 ; Franconni & Kamble, 2004].

Dans la majorité des propositions, nous trouvons les trois groupes d'opérations :

- Les opérations de **forage** permettent de naviguer au sein d'une hiérarchie de dimension pour analyser les indicateurs de performance avec plus ou moins de détails. Le forage vers le haut (*drill-up* ou *roll-up* ou *scale-up*) consiste à représenter les données d'un cube à un niveau de granularité supérieur conformément à la hiérarchie. Le forage vers le bas (*drill-down* ou *roll-down* ou *scale-down*) effectue l'opération inverse.

[7] http://business-intelligence.piloter.org/olap-evolution.htm

- Les opérations de **sélection** permettent de travailler sur une sous-partie des données d'un cube (*Slice* pour une restriction sur les attributs d'une dimension, *Dice* pour une sélection sur les mesures ou plus généralement *Select* dans certaines propositions).
- Les opérations de **rotation** permettent d'intervertir les axes d'analyse [Lehner, 1998] afin de présenter une autre face du cube. Pour une constellation, certains auteurs [Abello et al., 2003] proposent une rotation du sujet d'analyse (*Drill Across* ou *Frotate*).

Certains auteurs proposent de compléter ces opérations par les suivantes :

- **Modification du fait** : ces opérations permettent d'ajouter ou de supprimer une mesure dans un cube analysé [Cabibbo & Torlonne, 1998 ; Marcel, 1998]
- **Modification de la dimension** : ces opérations [Agrawal et al., 1997 ; Marcel, 1998] visent à transformer une mesure en attribut de dimension (*pull*) ou à effectuer l'opération inverse (*push*).
- **Ordonnancement** : ces opérations permettent soit d'ordonnancer les valeurs d'un attribut de dimension, soit l'emboîtement (*Nest*). L'emboîtement [Lehner 1998] consiste à imbriquer un paramètre sous un autre pour représenter sous une forme bi-dimensionnelle les informations d'un cube à N dimensions (avec N > 2).
- **Opérations ensemblistes** : certains auteurs proposent d'effectuer le produit cartésien, l'union, l'intersection ou la différence de cubes de données.

Ces langages de manipulation sont majoritairement définis à l'aide d'une algèbre [Agrawal et al., 1997 ; Cabibbo & Torlonne, 1997 ; Cabibbo & Torlonne, 1998 ; Gyssen & Lakshmanan, 1997 ; Lehner, 1998 ; Pedersen & Jensen, 1999 ; Abello et al., 2003 ; Franconi & Kamble, 2004]. Ils peuvent être également définis à l'aide d'un langage déclaratif à base de calcul [Li & Wang, 1996 ; Cabibbo & Torlonne, 1997] voire de règles [Marcel, 1998 ; Mendelzon & Vaisman, 2000].

Ces opérations peuvent s'appliquer sur des tables relationnelles [Li & Wang, 1996], sur des classes d'objets [Lehner, 1998 ; Mendelzon & Vaisman, 2000] mais plus majoritairement sur des cubes ou tableaux à *N* dimensions [Li & Wang, 1996 ; Agrawal et al., 1997 ; Cabibbo & Torlonne, 1997 ; Gyssen & Lakshmanan, 1997 ; Cabibbo & Torlonne, 1998 ; Marcel, 1998 ; Mendelzon & Vaisman, 2000 ; Abello et al., 2003 ; Franconi & Kamble, 2004].

Ces propositions présentent l'inconvénient de ne pas se soucier de la manière dont sont restituées les données aux décideurs. En effet, une visualisation sous forme de cubes en N dimensions (N > 2) semble difficilement exploitable par les décideurs [Gyssens & Lakshmanan, 1997] et ce type de présentation occulte la hiérarchisation des dimensions. Ces propositions sont incomplètes car aucune ne propose l'intégralité des opérations listées précédemment. De plus, ces propositions reposent sur des modèles de données logiques spécifiques et ne proposent pas toujours des définitions précises et formelles des différents concepts. A l'heure actuelle, il n'existe pas de consensus sur la définition d'un noyau minimum complet offrant une algèbre d'interrogation multidimensionnelle, à l'instar de l'algèbre relationnelle qui offre un support complet et reconnu. Enfin, seuls [Cabibbo & Torlonne, 1998] proposent la combinaison d'une algèbre et d'un langage graphique ; cependant le langage graphique se limite aux opérations de base et n'est pas complet au regard de l'algèbre.

3.4 METHODES DE CONCEPTION ET DEVELOPPEMENT

A notre connaissance, il n'existe pas de méthode de conception de SAD complet (reposant sur un entrepôt de données et des magasins de données) tel que nous l'avons défini dans la section 1.3. La plupart des travaux proposent une méthode pour la conception d'un magasin multidimensionnel qui reste le composant essentiel d'un SAD. L'analyse des démarches de ces

méthodes, nous permet de les classer en 3 catégories : **ascendante** (basée sur les données sources), **descendante** (basée sur les besoins utilisateurs) et **mixte**.

3.4.1 Méthodes ascendantes

Les **méthodes ascendantes** utilisent les sources de données pour concevoir les schémas multidimensionnels. Ces méthodes considèrent que les informations pertinentes pour la prise de décision se trouvent dans les sources [List et al, 2002] et que l'évolutivité des besoins utilisateurs est difficile à gérer [Moody & Kortink 2000]. Les auteurs de [Cabibbo & Torlone, 1998 ; Cabibbo & Torlone, 2000 ; Golfarelli & Rizzi, 1998 ; Moody & Kortink, 2000] proposent une méthode de conception d'un schéma multidimensionnel à partir de schémas Entité-Association décrivant les sources. L'avantage de cette démarche est qu'elle exploite complètement la sémantique des sources. L'inconvénient majeur est que le périmètre des données sources, parfois large, peut requérir des ressources en temps et en hommes. Pour contourner cette limite, les auteurs de [Husemann et al., 2000] proposent aux décideurs de préciser les données sources pertinentes. Dans cette proposition, l'étude des dépendances fonctionnelles des données sources permet de construire le schéma multidimensionnel résultat. L'inconvénient de cette solution est que les utilisateurs doivent consulter les schémas sources alors qu'ils ne les maîtrisent pas toujours.

3.4.2 Méthodes descendantes

Dans les **méthodes descendantes**, les données des sources ne sont pas prises en compte car, seuls les besoins des décideurs sont nécessaires pour définir un schéma multidimensionnel. R. Kimball [Kimball & Ross, 2002 ; Kimball et al., 2005] propose une méthode de gestion de projets décisionnels sans proposer une démarche précise pour une conception multidimensionnelle. De plus, ces travaux ne proposent pas de solution pour une modélisation conceptuelle des données. Les auteurs de [Tsois et al., 2001] proposent de concevoir un schéma multidimensionnel à partir d'une liste de requêtes exprimant les besoins des décideurs. Ces travaux se centrent essentiellement sur la définition de dimensions multi-hiérarchisées. Les auteurs de [Prat & Akoka, 2002 ; Prat et al., 2006] proposent une méthode de conception basée sur les notations UML et qui intègre les trois niveaux d'abstraction. Au niveau conceptuel, le concepteur traduit les besoins des décideurs en un diagramme de classes UML. Ce dernier est enrichi et transformé pour obtenir un schéma multidimensionnel. Les inconvénients de cette démarche est que seuls les besoins utilisateurs sont pris en compte et que le schéma conceptuel produit peut s'avérer impossible à mettre en œuvre par l'indisponibilité des données sources.

3.4.3 Méthodes mixtes

Les **méthodes mixtes** combinent les deux démarches précédentes et essayent de combler les lacunes de chacune d'elles. Elle se décompose en trois phases : (1) l'analyse des besoins utilisateurs produit un ou plusieurs schémas idéaux, (2) l'étude des sources se caractérise par un ou plusieurs schémas candidats (3) la phase de confrontation permet de comparer les schémas idéaux et candidats pour aboutir à un schéma final. Si la méthode produit plusieurs schémas idéaux [Carneiro & Braymer, 2002] ou plusieurs schémas candidats [Bonifati et al, 2001 ; Cavero et al., 2001; Phipps & Davis, 2002], voire les deux [Soussi et al., 2005], la phase de confrontation peut s'avérer complexe et fastidieuse. Seul [Luján-Mora & Trujillo, 2003] propose la confrontation d'un seul schéma candidat avec un seul schéma idéal. Les étapes de la démarche ne sont pas toujours explicitées [Carneiro & Braymer, 2002 ; Luján-Mora & Trujillo, 2003].

Certains travaux présentent quelques caractéristiques spécifiques tels que la prise en compte des traitements ETL [Luján-Mora & Trujillo, 2003] ou la gestion de la capitalisation et de la réutilisation [Carneiro & Braymer, 2002]. Certaines propositions minimisent la phase d'analyse voire l'occultent [Cavero et al., 2001 ; Phipps & Davis, 2002 ; Carneiro & Braymer, 2002 ; Luján-Mora & Trujillo, 2003]. L'expression des besoins peut s'effectuer de différentes manières : cas d'utilisation [Luján-Mora & Trujillo, 2003, Bruckner et al., 2001b], tableaux multidimensionnels

[Soussi et al., 2005] voire requêtes permettant de valider les schémas candidats [Phipps & Davis, 2002]. Il faut remarquer que depuis quelques années, une attention particulière est portée sur les modèles de buts pour exprimer les besoins décisionnels. Plusieurs modèles de buts ont été proposés : GDI [Prakash & Gosain, 2003], i* [Giorgini et al., 2005 ; Mazon et al., 2005a] et MAP [Gam & Salinesi, 2006]. Cependant, ces travaux doivent être complétés afin de proposer une méthode globale pour le développement de SAD [Mazon et al., 2005b].

3.5 BILAN SUR LES MAGASINS DE DONNEES

Cette troisième section nous a permis de présenter les différents concepts et travaux inhérents aux magasins de données OLAP. Les définitions proposées par [Codd et al., 1993] et [Pendse, 2005] reposant essentiellement sur des considérations commerciales, nous avons donné une nouvelle définition des systèmes OLAP en faisant abstraction des outils du marché et en mettant en avant l'aspect multidimensionnel des données. L'étude des travaux existants sur les modèles et les langages de manipulation nous a permis de mettre en évidence un manque de standardisation [Rizzi et al., 2006]. Cet ensemble de propositions hétérogènes ne permet pas toujours une expression aisée des besoins des décideurs (formalisme graphique absent ou non adapté, omission du niveau conceptuel, concepts limités, faible intégration de contraintes sémantiques et temporelles, pas de proposition pour la personnalisation de schéma etc.). Aucune méthode ne propose une démarche précise pour le développement d'un SAD complet. Les propositions se centrent sur le module multidimensionnel avec notamment une démarche mixte répondant de la manière la plus adéquate aux besoins des SAD. Cependant, dans de nombreuses propositions, la phase d'analyse est occultée voire absente.

4 POSITIONNEMENT DE NOS RECHERCHES

Durant ces dernières années, mes travaux ont consisté en la proposition de composants méthodologiques pour le développement de systèmes complets d'aide à la décision ainsi que l'exploitation de ceux-ci. Plus précisément, nos propositions reposent sur des modèles conceptuels pour les ED ou les MD, des langages d'analyse multidimensionnelle et une méthode pour le développement de SAD.

4.1 MODELISATION DES ENTREPOTS

Le développement d'applications tournées vers la mise en valeur à des fins décisionnelles du patrimoine informationnel stocké et géré par les applications traditionnelles de gestion nécessite de nouveaux modèles de données [Codd et al., 1993 ; Kimball & Ross, 2002]. Pour ce faire, nous souhaitons proposer une modélisation des entrepôts permettant d'intégrer des données sources hétérogènes et conserver leurs évolutions au cours du temps.

A notre avis, cette modélisation n'a pas requis toute l'attention qu'elle nécessite. Les premiers travaux relatifs aux ED se sont centrés essentiellement sur des aspects techniques d'extraction des données sources pour alimenter et rafraîchir un entrepôt (principe des vues matérialisées – cf. sections 2.1.1 et 2.1.2). Or, il est nécessaire de proposer des concepts et des formalismes spécifiques pour modéliser des entrepôts gérant de manière optimale l'important volume des données sources hétérogènes nécessaires aux prises de décisions.

Pour répondre à ce besoin, nous avons investigué dans deux directions peu ou pas traitées dans la littérature : la gestion optimale de données factuelles et la prise en compte de documents sources. Pour une gestion optimale des données factuelles, nous souhaitons que ce modèle puisse sauvegarder aussi bien les valeurs courantes des données extraites que leurs versions antérieures éventuellement agrégées.

Comme indiqué dans [Tseng & Chou, 2006], les systèmes d'aide à la décision construits uniquement à partir de bases de données sources, n'exploitent que 20% des informations

disponibles. Les 80% restants sont stockés dans des documents non structurés ou semi-structurés qu'il est judicieux d'intégrer dans un entrepôt. Cependant, les travaux proposés se limitent à des documents dont la structure est spécifique (proche d'un schéma de BD [Faulstich et al., 1997, 1998]) et définie au préalable [Abiteboul et al., 2001, 2002]. Dès lors que la structure n'est pas définie au préalable ou que les règles de structuration ne sont pas rigoureuses, il est difficile d'extraire la structure logique. Certains travaux n'exploitent pas la structure logique mais uniquement les liens hypertextes entre documents [Bhowmick et al, 2000a, 2000b]. De plus, tous les travaux relatifs à l'intégration de documents dans un SGBD traitent de la problématique de centralisation mais ne prennent pas en considération les analyses décisionnelles. Aussi, notre objectif est de proposer un modèle permettant :

- de spécifier des processus d'alimentation pour l'intégration de documents sources de différentes natures : structuré, semi-structuré voire non structuré,
- de générer automatiquement des classes de documents ayant des caractéristiques structurelles communes sans définition *a priori* de structure par un utilisateur,
- de stocker toutes les informations issues des documents sources (contenu, structures méta-données) pour effectuer des analyses décisionnelles de tous types : recherche d'information, interrogation déclarative, analyse multidimensionnelle.

4.2 MODELISATION DES MAGASINS DE DONNEES MULTIDIMENSIONNELLES

En l'état actuel, les propositions de modèles multidimensionnels sont effectuées à différents niveaux d'abstractions et ne supportent que partiellement l'ensemble des concepts multidimensionnels. De plus, comme indiqué dans [Rizzi et al., 2006] et [Niemi, et al. 2003], la modélisation OLAP des données ne repose pas sur une formalisation précise, stable et reconnue par l'ensemble de la communauté scientifique.

En réponse à ces différentes lacunes, nos travaux visent à proposer un modèle conceptuel orienté décideur. Cette orientation décideur a pour objectif d'offrir :

- une représentation proche des besoins des décideurs faisant abstraction de toute implantation physique,
- une modélisation reflétant l'ensemble des besoins décisionnels tels que la multiplicité des sujets d'analyse facilitant l'étude de corrélations et la multi-hiérarchisation des axes d'analyse,
- une modélisation fiable des données décisionnelles afin de gérer la cohérence sémantique (pour effectuer des croisements et des agrégations significatifs pour les prises de décisions) et temporelle (intégration de l'évolution des besoins utilisateurs et des sources de données),
- un ensemble de mécanismes facilitant les prises de décisions tels que l'intégration et la confrontation de commentaires de décideurs et la personnalisation des données précisant les données les plus pertinentes.

4.3 MANIPULATION DE DONNEES MULTIDIMENSIONNELLES

Comme exposé en section 3.3, de nombreuses propositions ont été faites pour la manipulation de données multidimensionnelles. Ces différentes propositions reposent sur des modèles spécifiques et ne couvrent qu'en partie l'ensemble des opérations réalisables lors d'analyses décisionnelles. La majorité des propositions repose sur une algèbre permettant de combiner différents opérateurs en réponse à une requête complexe. Mais, à l'heure actuelle, il n'existe pas de consensus sur la définition d'un noyau minimum complet offrant une algèbre d'interrogation multidimensionnelle comme pour l'algèbre relationnelle.

Dans la lignée de nos propositions sur les modèles multidimensionnels, nous souhaitons proposer une solution orientée décideur pour exprimer les analyses décisionnelles. Ce choix nous permettra dans un premier temps d'offrir une **algèbre orientée utilisateur** [Abelló et al., 2003]. Cette algèbre doit supporter des structures de restitution adaptées aux prises de décision et offrir l'ensemble des primitives algébriques nécessaires à l'élaboration de requêtes d'analyse décisionnelle. De plus, cette orientation décideur doit se traduire par la proposition d'un langage graphique et d'un langage assertionnel. Une telle complémentarité des langages est rarement proposée dans la littérature [Cabibbo & Torlonne, 1998] alors qu'elle est indispensable dans un contexte décisionnel.

4.4 METHODE DE CONCEPTION

Comme indiqué en section 3.4, il n'y a pas à l'heure actuelle de méthode permettant de concevoir un SAD complet (entrepôt et magasins de données). De par la spécificité des SAD (intégration de données sources hétérogènes, modélisation multidimensionnelle des magasins, architecture variée…) il n'est pas possible d'utiliser les méthodes de conception de SI existantes. Les propositions actuelles se sont essentiellement concentrées sur la modélisation des données multidimensionnelles. Or, les concepteurs décisionnels recherchent une méthode plus globale pour l'analyse, la conception et le développement d'un SAD complet.

En réponse à cette demande, notre objectif est de proposer une méthode comprenant :

- des concepts et des formalismes pour la modélisation des différents espaces de stockage,
- une démarche explicitant les différentes étapes à suivre tout en tenant compte aussi bien des besoins décideurs que des données sources,
- une phase d'analyse clairement identifiée,
- un outil d'aide à la conception.

4.5 PRESENTATION DE NOS RESULTATS

La suite de ce mémoire est décomposée comme suit. Le chapitre suivant expose nos solutions pour la modélisation des ED comportant aussi bien des données factuelles que des documents. Le troisième chapitre présente les travaux que nous avons réalisés pour la modélisation des magasins de données OLAP. Le quatrième chapitre se centre sur la manipulation de données multidimensionnelles décrite au travers d'une algèbre orientée utilisateurs et des langages graphique et assertionnel associés. Le cinquième chapitre décrit les étapes que nous avons définies pour la conception et le développement d'un système d'aide à la décision. Le sixième chapitre décrit les différentes thématiques abordées dans le cadre de nos travaux de recherche ainsi que les projets et productions scientifiques. Enfin, le dernier chapitre de ce mémoire propose un bilan de nos travaux et un ensemble de perspectives de travail.

CHAPITRE II : MODELISATION D'ENTREPOTS

PLAN DU CHAPITRE

1 INTRODUCTION A LA MODELISATION D'ENTREPOTS 31

2 ENTREPOT DE DONNEES EVOLUTIVES 32
2.1 Problématique 32
2.2 Concepts 33
2.2.1 Objet entrepôt 33
2.2.2 Classe entrepôt 34
2.2.3 Environnement et entrepôt 37
2.3 Formalismes et exemples 37

3 ENTREPOTS DE DOCUMENTS 39
3.1 Problématique 39
3.2 Modèle générique d'entrepôt de documents 40
3.2.1 Caractéristiques générales 40
3.2.2 Composants du modèle générique 40
3.3 Processus d'alimentation 41
3.4 Phase d'extraction 43
3.5 Phase de comparaison 44
3.5.1 Les étapes de la phase de comparaison 44
3.5.2 Exemple d'application 46

4 BILAN ET PERSPECTIVES 47
4.1 Bilan sur la modélisation des entrepôts 47
4.2 Production scientifique 48
4.3 Perspectives 49

1 INTRODUCTION A LA MODELISATION D'ENTREPOTS

Comme exposé dans le chapitre précédent, un Système d'Aide à la Décision (SAD) assure deux fonctions principales :

- centralisation, stockage et historisation des données sources utiles pour les décideurs,
- définition de sous-ensembles de données autour d'un thème particulier afin de répondre aux besoins spécifiques de ses utilisateurs.

Chacune de ces problématiques est traitée au travers d'espaces de stockage différents. Aussi, l'architecture des systèmes décisionnels que nous avons définie dès le début de nos travaux est basée sur une dichotomie d'espaces de stockage : l'entrepôt et les magasins de données [Teste, 2000]. Ce chapitre vise à apporter des solutions pour la modélisation des entrepôts.

Par essence, un entrepôt centralise des données sources disparates et hétérogènes. Cette hétérogénéité peut se traduire à différents niveaux : systèmes, modèles, formats et sémantiques des données [Kedad & Métais, 1999]. Une première solution pourrait consister à adapter les principes de la répartition des bases de données telle que je l'ai abordée dans mes travaux de thèse [Ravat, 1996], ou plus particulièrement, le principe des BD fédérées [Sheth & Larson, 1990]. Une BD fédérée repose sur l'intégration de bases de données sources (schéma et instances) hétérogènes pour produire une description unifiée des schémas initiaux (schéma intégré) et les règles de traduction [Sheth & Larson, 1990 ; Parent & Spaccapietra, 1996]. Or, cette solution n'est pas satisfaisante pour les raisons suivantes :

- par opposition aux BD fédérées, les entrepôts ne reposent pas une vision unifiée des sources mais permettent de stocker une agrégation des données sources hétérogènes avec leurs différentes évolutions dans le temps ;
- les ED construits uniquement à partir de BD sources n'extraient que 20% des informations disponibles [Tseng & Chou, 2006]. Sachant que les 80% restants sont stockés dans des documents [Tseng & Chou, 2006] il est préjudiciable pour un système décisionnel de ne vouloir intégrer que les données issues de BD sources.

Ces limites ont constitué la ligne directrice de mes travaux.

Dans un premier temps, nous proposons une solution adaptée pour la modélisation d'entrepôts afin de gérer de manière optimale l'ensemble des données factuelles sources nécessaires aux prises de décisions. De plus, comme indiqué dans le chapitre précédent, la prise en compte dans les ED de l'évolution des données au cours du temps est vitale pour les décideurs [Inmon, 1996 ; Chaudhuri & Dayal, 1997 ; Yang & Widom, 1998 ; Pedersen & Jensen, 1999 ; Yang & Widom, 2000]. Aussi, cette gestion optimale de données évolutives doit se traduire par une modélisation dans laquelle il est conservé :

- la valeur calculée à partir de la dernière extraction des informations sources,
- les éventuelles valeurs antécédentes nécessaires pour analyser les évolutions dans le temps,
- voire, les valeurs agrégées au bout d'un certain temps et dont le détail n'est pas utile pour les prises de décision.

Cette modélisation doit être accompagnée de la sauvegarde des processus d'alimentation intégrant les fonctions de calcul et de transformation des données sources.

Les données factuelles extraites des bases de production, ne peuvent constituer la seule source d'alimentation d'un entrepôt. Aussi, notre second objectif est de pouvoir intégrer dans un

entrepôt les contenus informationnel et structurel des documents nécessaires à la prise de décision [Fondin, 1998]. Pour ce faire, nous avons collaboré avec d'autres membres de l'équipe SIG qui possèdent une forte expertise en Système de Recherche d'Information (SRI). Plus précisément, nous avons souhaité combiner les SRI avec les SGBD pour définir un entrepôt de documents issus de sources disséminées. Contrairement à certaines propositions, nous souhaitons intégrer des documents structurellement hétérogènes sans connaissance *a priori* de leurs structures et sans imposer un langage spécifique. Cet entrepôt de documents sert de support à divers processus d'exploitation décisionnels dont l'objectif est triple : (1) rechercher des documents par leur contenu sémantique (type recherche d'information), (2) interroger l'entrepôt en mêlant à la fois les aspects structure et contenu (en utilisant les langages déclaratifs), et (3) enfin réaliser des analyses en se basant sur la structure au travers d'analyses multidimensionnelles.

Chacun de ces travaux est étudié dans les sections suivantes. La section 2 traite de la gestion optimale des données dans un entrepôt de données factuelles. La section 3 permet d'étudier l'intégration de documents dans un tel espace de stockage. La section 4 dresse un bilan des propositions et définit différentes perspectives.

2 ENTREPOT DE DONNEES EVOLUTIVES

Cette première partie de mes travaux vise à proposer des concepts et des formalismes pour la modélisation des entrepôts de données évolutives. Une première section vise à préciser la problématique de nos recherches et les suivantes présentent les résultats que nous avons obtenus.

2.1 PROBLEMATIQUE

Nous proposons de décliner l'objectif global qui est de modéliser les données décisionnelles évolutives au sein d'un ED en cinq points :

1. **Gestion efficace des données décisionnelles**. Un ED permet de stocker l'ensemble des informations décisionnelles utiles aux décideurs et ses évolutions au cours du temps. Un ED, également qualifié d'espace de préparation des données décisionnelles, est rarement manipulé en direct par les décideurs. Aussi, afin de garantir une gestion efficace des données et la pérennité de l'entrepôt, le modèle doit limiter la redondance tout en maintenant la cohérence des données. Le modèle proposé ne sera donc pas sous un format multidimensionnel.

2. **Gestion de données complexes**. Les données sources étant hétérogènes, le modèle doit offrir des abstractions suffisamment puissantes pour représenter l'intégration de ces différentes sources. De plus, la validation de ces travaux s'est effectuée dans un contexte médical qui utilise couramment des structures de données complexes [Lapujade & Ravat, 1997 ; Pedersen & Jensen, 1998 ; Pedersen & Jensen, 1999]

3. **Gestion de l'origine des données et extraction**. Afin de garantir la pérennité d'un ED, le modèle doit permettre de représenter les processus d'alimentation de l'ED présents et éventuellement passés en cas de modification de données sources ou de besoins décisionnels [Widom, 1995 ; Gupta & Mumick, 1995].

4. **Gestion de l'historique des données**. Les bases de production, orientées applications, ne sont pas prévues pour conserver l'historique des données ou ne conservent les évolutions que sur des périodes de temps insuffisantes pour les processus d'analyse décisionnelle. Le modèle de l'entrepôt doit donc être muni de mécanismes évolués permettant de représenter l'évolution des valeurs d'une donnée au cours du temps.

5. **Mécanisme d'archivage**. Une difficulté importante, liée à la conservation des évolutions des données, est le volume des données engendré par le stockage des historiques [Inmon, 1996]. De plus, au-delà d'une certaine date, les décideurs n'ont généralement pas besoin du détail de

toutes les évolutions. Notre modèle doit donc offrir la possibilité de conservation de données sous une forme archivée ou synthétique.

Comme indiqué dans le chapitre précédent, les travaux liés aux ED reposant sur le concept de vues matérialisées se soucient essentiellement des processus d'alimentation d'un ED stockant les données dans un format relationnel [Gupta & Mumick, 1995 ; Quass et al., 1996 ; Hull & Zhou, 1996 ; Zhou et al., 1996 ; Hurtado et al., 1999 ; Huyn, 1996 ; Huyn, 1997 ; Labio et al., 1999 ; Labio & Garcia-Molina, 1996 ; Mumick et al., 1997 ; Quass & Widom, 1997 ; Zhuge et al., 1995 ; Zhuge et al., 1996 ; Zhuge et al., 1997 ; Zhuge et al., 1998 ; Zhuge & Garcia-Molina, 1998 ; Kotidis & Roussopoulos 1999 ; Labio et al., 1997 ; Theodoratos & Sellis 1997 ; Theodoratos & Sellis 1999 ; Yang et al., 1997]. Certains travaux proposent le concept de vue matérialisée temporelle [Yang & Widom, 1998 ; Yang & Widom, 2000 ; Gupta & Mumick, 1999] mais ne proposent pas de solution pour une conception de données historisées et agrégées. A notre connaissance, il n'y a pas travaux proposant une modélisation adaptée des ED composés d'un large volume de données dont il faut conserver l'évolution dans le temps. Notre objectif est donc de proposer un modèle conceptuel pour répondre à ce besoin.

2.2 CONCEPTS

Même si l'approche objet [Cattell, 1998] confère une puissance d'expressivité et une richesse sémantique, elle n'est pas suffisante pour répondre entièrement à notre problématique. Aussi, nous avons étendu le concept d'objet pour aboutir à la définition d'un modèle conceptuel de données orienté objet intégrant des données temporelles et archivées [Ravat & Teste, 2000c].

Dans les sections suivantes, nous présentons les différents concepts de ce modèle. Vous trouverez une description détaillée de ceux-ci dans la thèse d'Olivier Teste [Teste, 2000].

2.2.1 Objet entrepôt

Le concept de base est l'**objet entrepôt**, extension du concept d'objet traditionnel pour modéliser l'aspect évolutif des données. Pour ce faire, nous avons associé à chaque objet entrepôt trois types d'états différents : un **état courant**, ses **états passés** et un résumé de ses états passés successifs agrégés au travers d'**états archivés**.

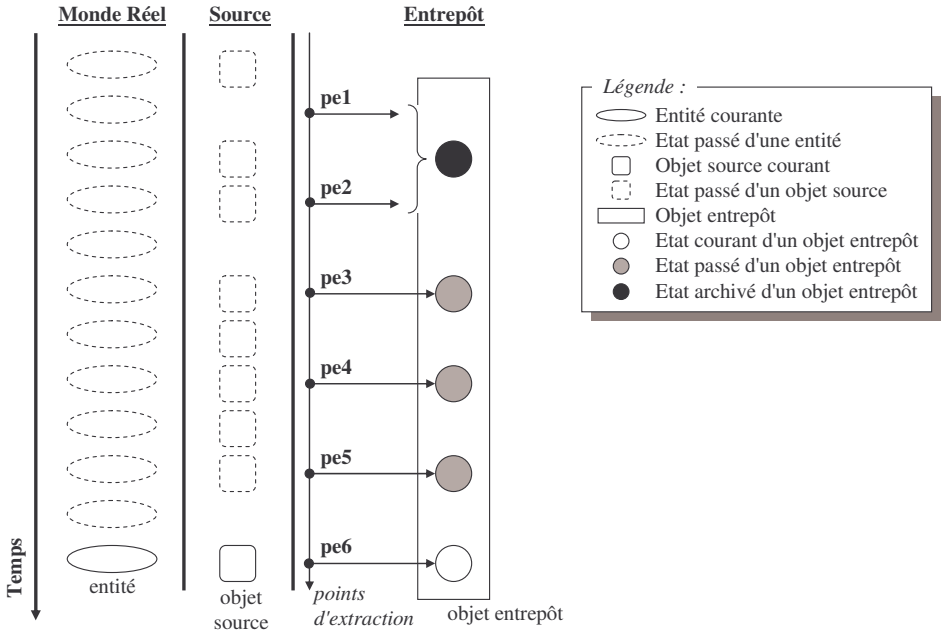


Figure 7 : Modélisation d'un objet entrepôt

La Figure 7 illustre notre proposition en représentant un objet entrepôt composé d'un état courant, de trois états passés et d'un état archivé construit à partir de deux anciens états passés qui sont supprimés.

Nous définissons le concept d'objet entrepôt par l'expression suivante :

Définition : Un **objet entrepôt** o est défini par un n-uplet (oid, S_0, $Histoire$, $Archive$, $Origine$) :

- oid est l'identifiant interne,
- S_0 est l'état courant (dernières valeurs extraites pour les propriétés de l'objet o),
- $Histoire = \{S_{p1}, S_{p2}, \ldots, S_{pp}\}$ est l'ensemble fini d'états passés correspondant au détail des évolutions de valeur des propriétés déclarées comme temporelles,
- $Archive = \{S_{a1}, S_{a2}, \ldots, S_{aa}\}$ est l'ensemble fini d'états archivés correspondant à une agrégation de certaines évolutions détaillées (états passés),
- $Origine = \{s\_oid1, s\_oid2, \ldots, s\_oids\}$ est l'ensemble fini des identifiants des objets source à partir desquels la valeur de l'objet entrepôt est obtenue. Lorsqu'un objet source est supprimé, l'objet entrepôt associé ne peut plus être rafraîchi : il est figé.

Nous définissons le concept d'état par l'expression suivante :

Définition : Un **état** S_i se définit par un couple ($IntT_i$, V_i) où

- $IntT_i = [DateDeb, DateFin]$ est un **intervalle temporel** correspondant aux instants durant lesquels l'état S_i a été courant,
- V_i est **valeur structurelle** associée à la **valeur temporelle** ($DomT_i$), autrement dit, la valeur des propriétés de l'objet durant les instants de $IntT_i$.

2.2.2 Classe entrepôt

De manière classique en objet, les objets entrepôt ayant une même structure et un même comportement sont regroupés dans une classe. Nous avons étendu le concept standard de classe [Cattell, 1998] afin d'y ajouter les processus d'extraction (correspondance entre les objets entrepôt et leurs origines) ainsi que les propriétés temporelles et archivées.

Définition : Une **classe entrepôt** c est définie par un n-uplet (Nom^c, $Type^c$, $Super^c$, $Extension^c$, $Mapping^c$, $Tempo^c$, $Archi^c$) où

- Nom^c est le nom unique de la classe,
- $Type^c$ est le type de la classe, autrement dit, le schéma précisant la structure et le comportement communs aux objets,
- $Super^c$ est l'ensemble des super-classes de c,
- $Extension^c = \{o_1, o_2, \ldots, o_x\}$ est l'ensemble fini d'objets entrepôt de la classe c,
- $Mapping^c$ est la **fonction de dérivation** qui caractérise le processus d'extraction à partir duquel la classe c est générée et alimentée en instances,
- $Tempo^c$ est le **filtre temporel** qui liste les attributs dont les évolutions de valeur seront conservées à chaque point d'extraction (rafraîchissement de la classe entrepôt),
- $Archi^c$ est le **filtre d'archivage** qui liste les attributs pour lesquels l'évolution de valeurs peut être stockée de manière agrégée.

Dans les sections suivantes, nous étudions les caractéristiques spécifiques aux classes entrepôt, à savoir, les filtres temporel et d'archivage ainsi que la fonction de dérivation.

2.2.2.1 Les filtres

Les filtres permettent de préciser la gestion des évolutions de valeurs de données d'une classe entrepôt. Plus précisément, le **filtre temporel** définit la structure des états passés en listant des propriétés de la classe pour lesquelles les valeurs sont conservées à chaque rafraîchissement de l'entrepôt. Le **filtre d'archivage** précise la manière dont sont agrégées les données temporelles. Il est défini comme suit : $Archᵉ=\{(a_1, f_1), (a_2, f_2),\ldots, (a_n, f_n)\}$ où quel que soit (a_j, f_j), a_j est un attribut appartenant à $Tempoᵉ$ et f_j est une fonction d'agrégation précisant la manière dont sont résumées les valeurs temporelles de a_j.

Dans un filtre d'archivage, chaque attribut est associé à une fonction d'agrégation. Nous avons proposé deux types d'agrégation :

- Agrégation classique (*sum*, *count*, *avg*, *max*, *min*) pour résumer l'ensemble des états passés dans un seul état archivé,
- Agrégation temporelle (*t_sum(T), t_avg(T), t_count(T)*…) pour résumer les états passés dans plusieurs états archivés, chacun correspondant à une période de temps *T*.

Une classe repose sur un **archivage fort** si son filtre d'archivage comporte des fonctions d'agrégations classiques sinon elle repose sur un **archive modéré**.

Exemple : La figure suivante illustre les deux types d'archivages. La Figure 8(a) représente l'archivage fort d'un objet entrepôt. La Figure 8(b) présente un archivage modéré permettant de regrouper par période de 5 ans les états passés de l'objet.

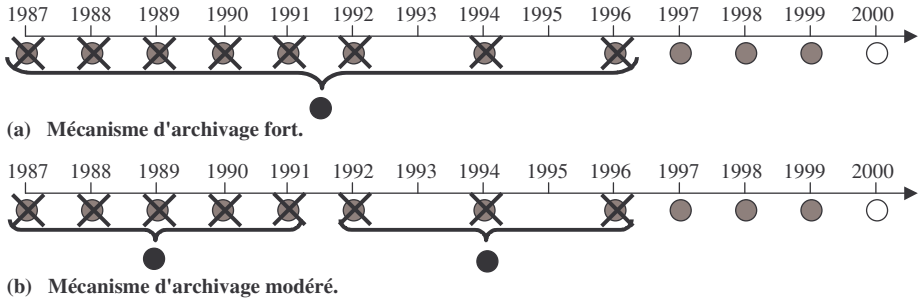


Figure 8 : Principes du mécanisme d'archivage

2.2.2.2 Fonction de dérivation

Cette fonction de dérivation permet de modéliser le processus de construction des classes entrepôt à partir de données sources. Pour répondre à ce besoin, nous avons proposé une algèbre. Dans une première section, je vous présente le cadre de nos travaux et dans la seconde, les opérateurs algébriques.

Sachant que les données sources d'un ED sont hétérogènes, nous avons subdivisé la partie extraction des données en deux sous-parties. La première, l'**intégration** se propose de résoudre les problèmes d'hétérogénéité (systèmes, modèles, formats et sémantiques des données,…) [Kedad & Métais, 1999] des différentes sources de données en intégrant celles-ci dans une source globale. Cette source globale est virtuelle, c'est à dire que les données utilisées pour la décision restent stockées dans les sources de données et sont extraites uniquement au moment des mises à jour de l'entrepôt. La source globale est décrite au moyen du modèle de données orientées objet car il s'avère parfaitement adapté pour l'intégration de sources hétérogènes [Bukhres & Elmagarmid, 1995]. La **construction** consiste à extraire de la source globale les données pertinentes pour la prise de décision, puis à les recopier dans l'entrepôt de données.

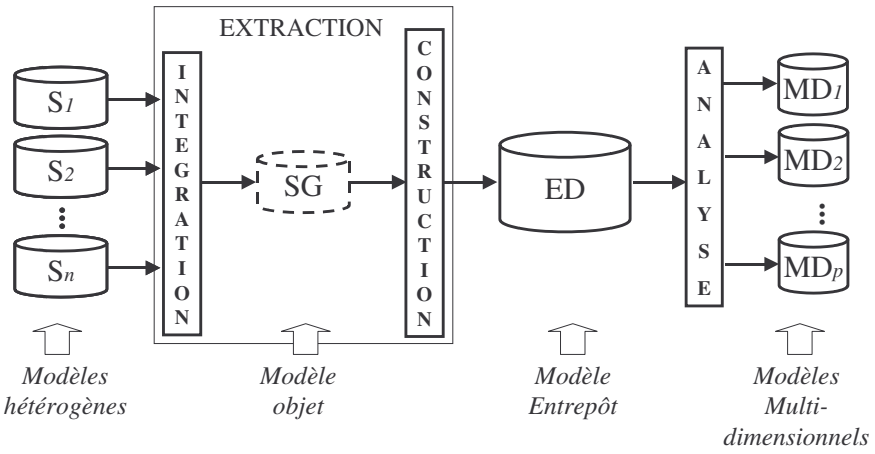


Figure 9 : Les composants du processus d'extraction

La fonction de dérivation (*Mapping*[c]) décrit le **processus de construction des structures** des classes entrepôt à partir d'une SG. Pour la définition de cette fonction, nous avons proposé une algèbre comprenant des opérations de qualification, de structuration, ensemblistes et de hiérarchisation. Le tableau suivant donne la syntaxe et la description de ces différentes opérations. Un exemple complet d'utilisation de cette algèbre est disponible dans [Teste, 2000].

CATEGORIES	OPERATIONS	DESCRIPTIONS	
Qualification	**Sélection**	Syntaxe :	$\sigma(cs, pred)=c$
		Rôle :	Restreindre l'extension de cs à partir de laquelle est peuplée celle de c.
	Jointure	Syntaxe :	$\bowtie(cs_1, cs_2, pred\_j)=c$
		Rôle :	Filtrer le produit cartésien entre cs_1 et cs_2 à partir duquel est peuplée l'extension de c.
	Groupement	Syntaxe :	$\eta(cs, \{p_1, p_2,\ldots,p_m\}, attr)=c$
		Rôle :	Regrouper des objets de cs pour peupler c.
	Dégroupement	Syntaxe :	$\delta(cs, attr)=c$
		Rôle :	Décomposer des objets multivalués de cs pour peupler c.
Ensemblistes	**Union**	Syntaxe :	$\cup(cs_1, cs_2)=c$ ou $\cup^V(cs_1, cs_2)=c$
		Rôle :	Réaliser l'union de cs_1 et cs_2 pour générer c.
	Intersection	Syntaxe :	$\cap(cs_1, cs_2)=c$ ou $\cap^V(cs_1, cs_2)=c$
		Rôle :	Réaliser l'intersection de cs_1 et cs_2 pour générer c.
	Différence	Syntaxe :	$-(cs_1, cs_2)=c$ ou $-^V(cs_1, cs_2)=c$
		Rôle :	Réaliser la différence entre cs_1 et cs_2 pour générer c.
Structuration	**Projection**	Syntaxe :	$\pi(cs, \{p_1, p_2,\ldots,p_m\})=c$
		Rôle :	Définir les propriétés de cs qui sont extraites pour élaborer la structure de c.
	Masquage	Syntaxe :	$\mu(cs, \{p_1, p_2,\ldots,p_m\})=c$
		Rôle :	Définir les propriétés de cs qui ne sont pas extraites pour élaborer la structure de c.
	Accroissement	Syntaxe :	$\alpha(cs, \{p_1:e_1, p_2:e_2,\ldots,p_m:e_m\})=c$
		Rôle :	Définir les propriétés supplémentaires qui sont ajoutées à la structure de c élaborée avec l'ensemble des propriétés de cs.
Hiérarchisation	**Généralisation**	Syntaxe :	$\Lambda(c_1, c_2,\ldots, c_n)=c$
		Rôle :	Créer une super classe à $c_1, c_2,\ldots, c_n$ dont le type est constitué des propriétés communes.
	Spécialisation	Syntaxe :	$\Sigma(c_1, c_2,\ldots, c_n, pred\_j)=c$
		Rôle :	Créer une sous classe à $c_1, c_2,\ldots, c_n$.

2.2.3 Environnement et entrepôt

Afin de ne conserver que les évolutions pertinentes, nous souhaitons offrir la possibilité de définir au sein d'un ED une ou plusieurs parties temporelles dont la taille est en adéquation avec les exigences des applications décisionnelles. Chacune de ces parties temporelles peuvent être rafraichie selon des périodicités spécifiques (**hétérogénéité du rafraichissement**). Ces parties temporelles peuvent supporter différentes **granularités d'historisation** : un attribut pour les différents objets d'une classe (granularité attribut), les objets eux-mêmes (granularité classe) ou plusieurs classes et leurs relations (granularité ensemble).

Pour répondre à ce besoin, nous proposons le concept d'**environnement** [Ravat et al., 1999], permettant de définir des parties temporelles homogènes, cohérentes et configurables.

Définition : Un **environnement** *env* est défini par (Nom^{env}, C^{env}, $Config^{env}$) où

- Nom^{env} est le nom de l'environnement,
- $C^{env}=\{c_{e1}, c_{e2},\ldots,c_{en}\}$ est l'ensemble fini des classes contenues dans l'environnement,
- $Config^{env}$ détermine le comportement temporel local de l'environnement (période de rafraîchissement et critère d'archivage des états passés des classes) au travers d'un langage de configuration basé sur l'approche ECA (Evènement-Condition-Action) [Dayal, et al., 1988 ; Widom, 1996].

Le concept d'environnement permet d'unifier l'historisation en offrant trois niveaux de **granularité d'historisation** :

- **Granularité attribut** : Cette granularité, la plus fine, consiste à conserver les évolutions d'un ou plusieurs attributs d'une seule classe. Pour cela, l'administrateur définit un environnement englobant uniquement la classe ; il définit le filtre temporel de la classe contenant seulement les attributs à historiser.
- **Granularité classe** : La granularité classe consiste à conserver les évolutions de chaque objet pour l'ensemble des valeurs des attributs de la classe. Pour ce faire, l'administrateur doit définir un environnement englobant uniquement la classe à historiser. Son filtre temporel contient l'ensemble des attributs de la classe, garantissant ainsi la conservation de leurs évolutions détaillées au travers d'états passés.
- **Granularité ensemble** : Cette granularité vise à conserver les évolutions de plusieurs classes d'objets entrepôt interconnectées par des relations d'association et/ou de composition. Le filtre temporel de chacune des classes historisées contient l'ensemble des attributs et des relations dont les évolutions détaillées sont conservées.

Pour clore cette présentation de notre modèle, il reste à spécifier le concept d'entrepôt :

Définition : Un **entrepôt** *ED* est défini par le n-uplet (Nom^{ED}, C^{ED}, Env^{ED})

- Nom^{ED} est le nom de l'entrepôt,
- $C^{ED}=\{c_1, c_2,\ldots, c_n\}$ est l'ensemble fini des classes de l'entrepôt,
- $Env^{ED}=\{env_1, env_2,\ldots, env_m\}$ est l'ensemble fini des environnements de l'entrepôt,

2.3 FORMALISMES ET EXEMPLES

Pour la description des différents composants d'un schéma d'ED, nous avons proposé un formalisme textuel et un formalisme graphique. Le formalisme textuel proposé est basé sur une extension du langage de définition standard de l'ODMG [Cattell, 1998] comprenant la

composition d'objet [Bertino & Guerrini, 1998] ainsi que la définition des environnements, des filtres temporels et des filtres d'archivage. Tous les détails sont disponibles dans [Teste, 2000].

Pour faciliter la tâche du concepteur, nous avons également proposé un formalisme graphique. Extension du diagramme de classe UML[8], ce formalisme possède les caractéristiques suivantes :

- taxonomie des propriétés par ajout d'un préfixe ('D_' pour une propriété dérivée, 'S_' pour un élément spécifique et 'C_' pour un attribut calculé),
- ajout d'un pictogramme pour les attributs d'un filtre temporel ( ) ou d'archivages ( ).

Exemple : Le concepteur décisionnel souhaite définir un entrepôt de données permettant de conserver les informations relatives aux praticiens hospitaliers. Plus précisément, ces praticiens exercent dans les services d'un établissement hospitalier. Le décideur souhaite conserver les évolutions détaillées des dépenses de fonctionnement des différents services ainsi que les dépenses totales des établissements. De plus, le décideur souhaite archiver les dépenses des établissements de manière bi-annuelle (tous les deux ans).

Pour répondre à ce besoin, le concepteur doit définir 4 classes entrepôt : *PERSONNES, PRATICIENS, SERVICES, ETABLISSEMENTS*. Pour la conservation des évolutions de valeurs, il doit définir un environnement contenant les classes *ETABLISSEMENTS* et *SERVICES* et pour chacune d'elles, il doit définir des filtres temporels et d'archives.

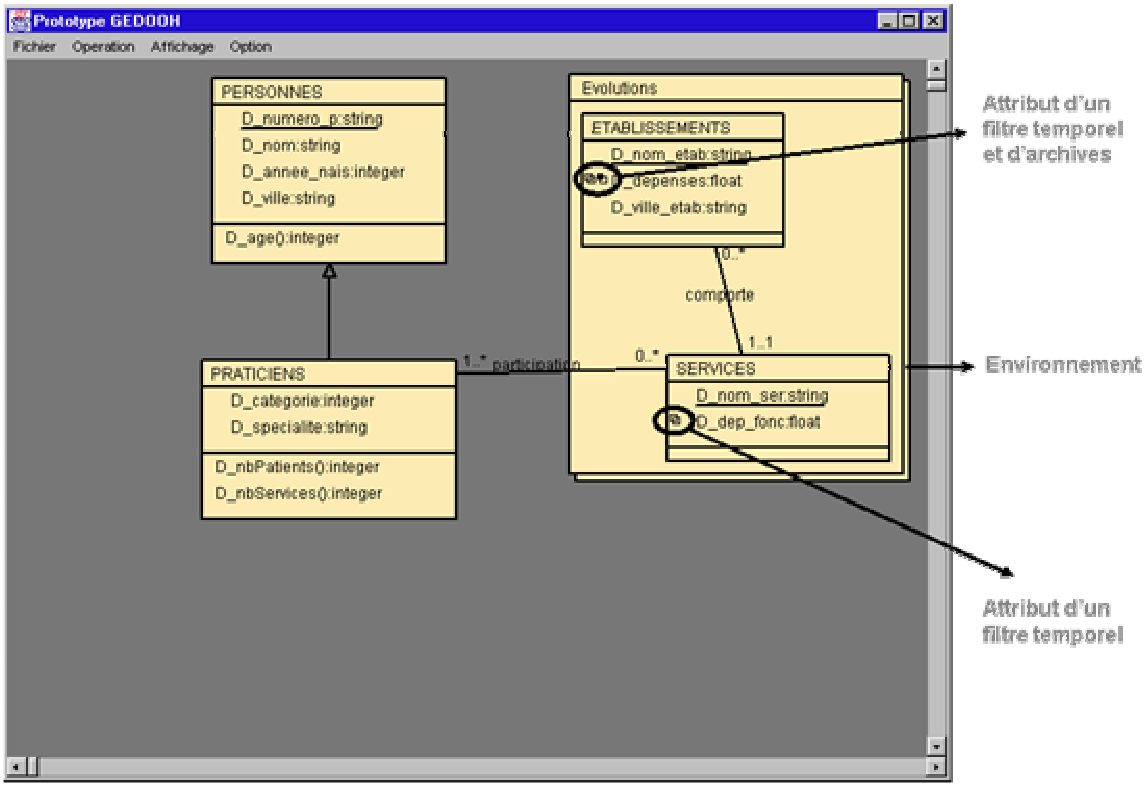


Figure 10 : Exemple d'un formalisme graphique

[8] http://www.uml.org/

3 ENTREPOTS DE DOCUMENTS

La prise en compte de l'hétérogénéité des sources d'un SAD consiste à intégrer non seulement des données factuelles mais également les documents utiles pour un processus de prise de décision. L'objectif de cette section est de présenter les résultats que nous avons obtenus pour la spécification d'entrepôts à base de documents. Dans une première section, nous explicitons la problématique de recherche. Dans les sections suivantes, nous étudions les concepts et le processus d'alimentation du modèle d'entrepôt de documents que nous proposons.

3.1 PROBLEMATIQUE

Tel que défini dans le premier chapitre, un **Entrepôt de DOcuments textuels (EDO)** est un espace de stockage centralisé d'informations issues des documents sources pertinents pour les décideurs. Un tel entrepôt sert de support aux processus de recherche d'information, d'interrogation et d'analyse décisionnelles. Cette définition induit la résolution de deux problèmes.

Le premier problème est relatif à l'intégration et au stockage d'informations issues de documents hétérogènes. Cette hétérogénéité peut être aussi bien structurelle que sémantique. L'hétérogénéité sémantique se traduit par des contenus très divers et non limités *a priori*. L'hétérogénéité structurelle consiste à prendre en compte des documents possédant des structures logiques différentes voire des documents plus ou moins structurés (documents structurés, semi-structurés ou non structurés) et non nécessairement définis à l'aide de standard.

Le second problème consiste à déterminer les informations à extraire des documents sources et le format de stockage pour être disponibles aux différents processus de restitution. Notamment, nous souhaitons que tout EDO puisse supporter ces différentes techniques de restitutions complémentaires :

- la recherche d'informations pour restituer des "granules" ou des documents entiers en réponse à une requête utilisateur formulée à l'aide de mots-clés (et pallier les inconvénients des moteurs de recherche restituant des listes de références de documents contenant au moins un de ces mots-clés),
- l'interrogation de données pour restituer des données factuelles en réponse à une requête exprimée à l'aide d'un langage déclaratif et dont les prédicats de sélection portent sur des éléments de structure ou de contenu,
- l'analyse multidimensionnelle sur des informations documentaires (éléments de contenu et/ou de structure).

Les Systèmes de Recherche d'Information (SRI) traditionnels, basés sur l'indexation de termes, n'apportent pas de solutions satisfaisant nos besoins. Les SRI stockant les documents dans des fichiers, les processus d'alimentation de documents hétérogènes seraient fastidieux à développer et le processus de manipulation décisionnelle telle que nous venons de l'énoncer semble inapplicable. Un entrepôt de documents n'a pas vocation à stocker des documents bruts mais il doit les transformer pour les rendre plus adaptés aux applications utilisatrices. De part leur méthode de stockage et leur langage déclaratif d'interrogation, les SGBD sont mieux adaptés à une gestion de ce type d'informations documentaires [Soulé-Dupuy, 2001].

De nombreux travaux ont apporté des solutions pour le stockage dans un SGBD de la structure logique des documents structurés [Comparot, 1994 ; Comparot & Chrisment, 1994 ; Christophidès, 1996 ; Abiteboul, 1997 ; Abiteboul & Vianu, 1997 ; Abiteboul et al., 2002] ou semi-structurés [Gardarin & Yoon, 1996 ; McHugh et al., 1997 ; Adelberg, 1998 ; Nestorov et al., 1998 ; Riahi, 1998 ; Goldman et al., 1999 ; Mothe et al., 2000 ; Riahi, 2000]. Quant aux travaux relatifs aux entrepôts documentaires, ils se limitent à l'intégration de documents dont la structure logique est spécifique (proche d'un schéma de BD [Faulstich et al., 1997, 1998] ou exprimée à

l'aide de liens hypertextes [Bhowmick et al, 2000a, 2000b]) et définie au préalable [Abiteboul et al., 2001, 2002]). Cependant, la faiblesse de ces travaux est mise en évidence lorsque la structure logique est absente ou n'est pas spécifiée (textes bruts, textes numérisés) ou lorsque cette structure logique, par son manque de contraintes, permet d'aboutir à des documents trop hétérogènes au niveau de leurs structures. Ces limites nous offraient un nouveau champ d'investigation. Notre objectif est de développer un EDO comme une base d'informations synthétiques et homogènes issues de documents sources hétérogènes plus ou moins structurés. Cet EDO doit également faciliter son exploitation décisionnelle. Pour répondre à ce besoin, nous proposons un modèle générique et un processus d'alimentation associé.

3.2 MODELE GENERIQUE D'ENTREPOT DE DOCUMENTS

Le modèle générique que nous proposons permet de représenter le contenu, les méta-informations ainsi que les structures logiques génériques et spécifiques des documents sources. Toute structure logique est représentée à l'aide d'une arborescence ordonnée et étiquetée dans laquelle chaque nœud identifie un élément de la structure logique (identifié par une balise).

3.2.1 Caractéristiques générales

D'un point de vue stockage, la prise en compte des deux structures logiques permet :

- une indexation plus fine des documents à base de **granules** ou une unité d'information "sémantiquement homogène" d'un document,
- de regrouper les documents en collections structurellement homogènes.

D'un point de vue manipulation, la définition de granule permet des restitutions plus ciblées et donc plus pertinentes pour les décideurs. De plus, l'intégration du contenu, de méta-informations et des structures, permet des manipulations combinant ces différents composants.

Les entrepôts de documents conçus selon un modèle générique doivent être évolutifs et indépendants de toute norme de représentation. Ils doivent être également indépendants du niveau de granularité des documents pour permettre ainsi tout type d'exploitation de contenu [Soulé-Dupuy, 2001].

3.2.2 Composants du modèle générique

Nous avons représenté ce modèle générique à l'aide d'un diagramme de classes UML[9]. Pour en faciliter la compréhension, nous avons décomposé ce diagramme de classes en trois parties représentant la structure logique générique, la structure logique spécifique et le contenu.

Une structure logique générique est décrite comme suit :

- la **structure logique générique** est caractérisée par un nom et un ensemble d'éléments génériques du premier niveau ;
- un **élément générique** est un composant de la structure logique (DTD par exemple) d'un ou de plusieurs documents. Il s'agit d'un nœud de l'arborescence qui représente la structure logique. Un élément générique est caractérisé par un nom unique, une cardinalité et éventuellement des éléments ou des attributs génériques ;
- un **attribut générique** est associé à un élément générique pour le décrire, c'est à dire ajouter des informations concernant l'élément en question, comme par exemple : indiquer son rôle, faire appel à des données externes.

[9] http://www.uml.org/#UML2.0

Une structure logique spécifique à un document est une instance d'une structure logique générique. Elle est décrite au travers des composants suivants :

- une **structure logique spécifique** précise le nom du document, sa date de création, la structure générique associée, les éléments spécifiques de premier niveau, et un ensemble de déclaration ;
- un **élément spécifique** est un granule du document qui doit correspondre obligatoirement à un élément générique de la structure logique générique de son document. Un élément spécifique est caractérisé par un numéro de séquence (ou ordre d'apparition de l'élément spécifique pour l'élément générique considéré), l'élément générique correspondant, la liste des sous-éléments spécifiques de niveau immédiatement inférieur (fils), ses attributs spécifiques et l'information associée ;
- un **attribut spécifique** décrit une information concernant un attribut d'un élément spécifique. Cet attribut doit correspondre obligatoirement à un attribut générique.

Le contenu textuel est basé sur les techniques d'indexation automatique de textes. Il permet d'associer à chaque granule informationnel ses mots-clés et ses fréquences d'apparition utilisées dans le cadre d'une restitution basée sur la recherche d'information. Dans le cadre de nos travaux, nous utilisons les deux fréquences suivantes :

- la fréquence absolue "Fréq_Abs" qui correspond au nombre d'occurrences du terme dans la collection de documents de l'entrepôt,
- la fréquence relative "Fréq_Rel" qui correspond au nombre d'occurrences du terme dans un granule documentaire.

Vous retrouvez ces différents composants dans le diagramme de classes UML suivant :

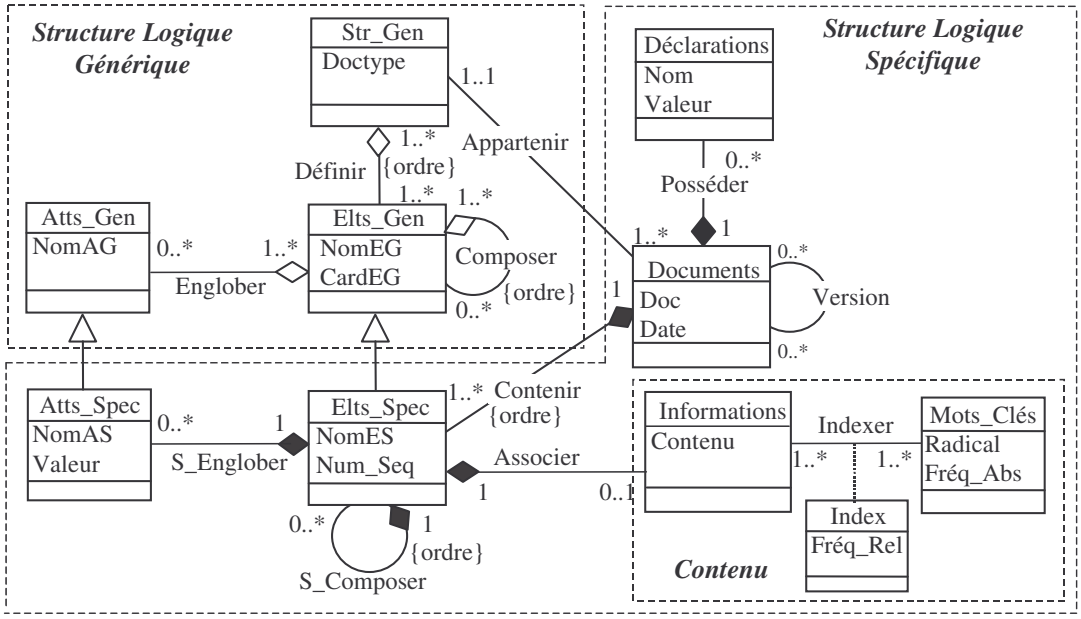


Figure 11 : Modèle générique d'entrepôts de documents textuels

Vous trouverez dans [Khrouf, 2004] des exemples d'instanciation de ce modèle à partir de documents structurés (XML valide), semi-structurés (HTML) et non structurés (TXT).

3.3 PROCESSUS D'ALIMENTATION

Une fois le modèle générique validé [Soulé-Dupuy, 2001 ; Khrouf, 2001c], une deuxième problématique s'offrait à nous ; il s'agissait de déterminer les processus d'alimentation d'un tel

entrepôt à partir de documents hétérogènes. Nous proposons de subdiviser cette problématique générale en trois phases :

- **Extraction de structure et de contenu** : cette phase doit permettre d'extraire la structure logique, ainsi que le contenu du document à insérer dans l'entrepôt [Khrouf & Soulé-Dupuy, 2003]. Afin d'intégrer des documents hétérogènes, il est nécessaire de définir une technique d'extraction pour chacun des trois types de documents que nous avons identifiés précédemment ;
- **Comparaison de structures** : cette phase consiste à comparer la structure logique extraite du document lors de la phase précédente avec toutes les structures logiques génériques existantes dans l'entrepôt [Khrouf & Soulé-Dupuy, 2003]. Si une structure identique ou approchante existe, le système *fusionne* les deux structures en une structure logique générique et rattache le document à cette structure. Dans le cas contraire, le système *crée* une nouvelle structure logique générique afin d'y rattacher le document ;
- **Insertion de contenu** : cette phase consiste à insérer le contenu du document dans l'entrepôt en rattachant chaque granule ou partie à l'élément de structure correspondant. Cette phase consiste en la définition d'un "parseur" (analyseur syntaxique) pour réaliser l'insertion.

Nous pouvons schématiser ce processus avec la figure suivante :

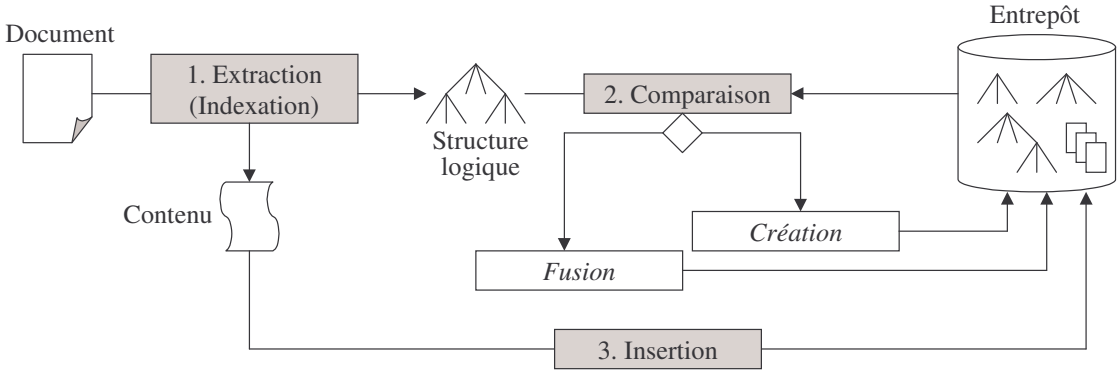


Figure 12 : Processus d'alimentation d'un entrepôt de documents

Nous illustrons ce processus au travers d'un exemple (Figure 13). Nous souhaitons intégrer le document "DocN.xml" dans un entrepôt contenant deux structures logiques génériques auxquelles sont rattachés deux ensembles disjoints de documents. Ce processus d'intégration doit suivre les 3 phases suivantes :

- La phase d'extraction a permis de spécifier la structure logique et les éléments de contenu du document "DocN.xml".
- La phase de comparaison a permis de définir que la structure logique de "DocN.xml" et la structure logique générique n°2 de l'entrepôt étaient suffisamment "proches" pour représenter une même classe de documents et pour pouvoir être fusionnées. Cette structure logique générique a été modifiée pour tenir compte des spécificités liées au nouveau document (ajout de l'élément "Résumé" et modification des cardinalités pour les éléments "Auteur" et "Section"), sans pour autant perdre les caractéristiques concernant des documents précédemment rattachés.
- Lors de la phase d'insertion, le contenu de "DocN.xml" a pu être rattaché aux différents éléments de la structure logique générique n°2 modifiée lors de la phase précédente.

La figure suivante représente ces trois phases :

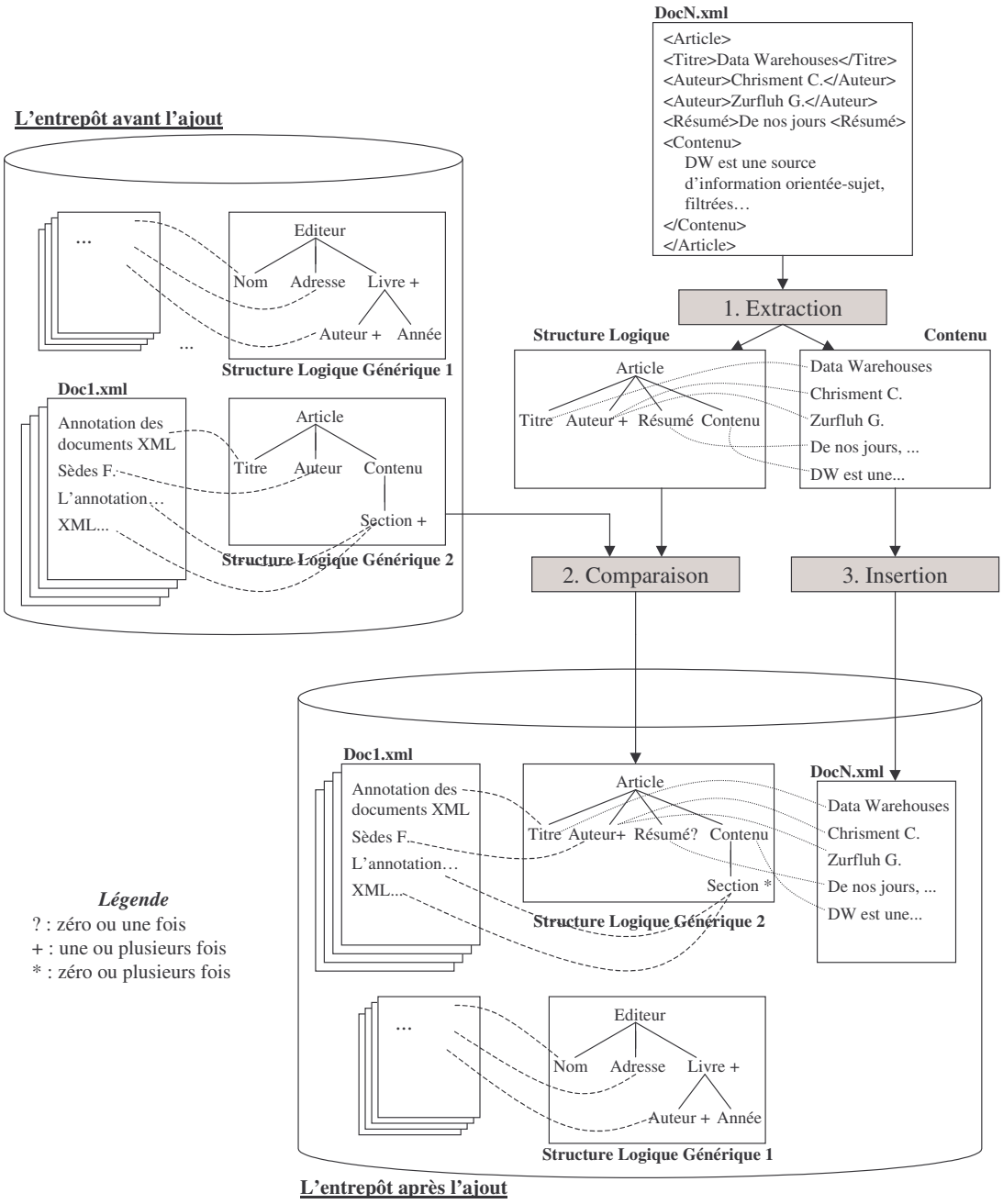


Figure 13 : Exemple d'alimentation d'un entrepôt

Dans les sections suivantes, nous présentons les problèmes que nous avons dû résoudre pour les phases d'extraction et de comparaison. La phase d'insertion reposant sur un "parseur" ne relève pas d'une difficulté majeure.

3.4 PHASE D'EXTRACTION

Cette phase permet de dissocier le contenu de la structure logique spécifique d'un document source à intégrer dans l'entrepôt.

Pour le contenu, nous proposons d'utiliser les techniques d'indexation [Khrouf, 2004] couramment utilisées en Recherche d'Information (RI). L'indexation permet d'associer à un document un ensemble d'informations le caractérisant. Ces informations peuvent être des mots-clés ou des méta-informations (auteur, résumé date d'édition etc.). Cette partie n'ayant pas fait

l'objet de développement spécifique de notre part, nous proposons d'utiliser les travaux définis pour l'indexation manuelle ou automatique [Rijsbergen, 1979 ; Salton & McGill, 1983 ; Salton, 1989].

L'étude de la structure logique spécifique d'un document source permet de construire une **arborescence ordonnée et étiquetée**. L'arborescence définit l'organisation des différents granules d'informations du document. Les étiquettes de l'arborescence correspondent aux balises de la structure logique du document. Chaque élément de cette arborescence peut être optionnel (cardinalité ?), monovalué ou multivalué (+ pour "1 ou plusieurs" et * pour "0 ou plusieurs").

Pour cette étape, la difficulté consiste à développer des algorithmes permettant d'interpréter et/ou éventuellement d'ajouter des balises pour chacun des 3 types de documents sources (structurés, semi-structurés ou non structurés). Ces balises vont constituer les étiquettes de l'arborescence représentant une structure logique.

Pour les documents structurés contenant des balises à "vocation sémantique", les algorithmes étaient simples à écrire. Pour les documents valides (reposant sur une DTD pour un document XML par exemple), la construction de l'arborescence est automatique. Pour les documents bien formés, l'algorithme consiste à ajouter d'éventuelles balises manquantes, mais l'algorithme reste simple également.

Pour les documents semi-structurés (comme par exemple les documents HTML), l'extraction de la structure s'avère plus difficile. L'extraction ne peut être automatique et nécessite une étape de réécriture. Cette étape de réécriture consiste à éliminer ou substituer les balises à vocation de présentation afin de mettre en évidence les éléments structurels et le contenu. Tous les détails de cette étape de réécriture se trouvent dans [Khrouf, 2004] et [Fualdes, 2001].

Pour les documents non structurés, nous proposons trois étapes :

- segmenter le document en paragraphes [Salton et al., 1997 ; Lallich & Ouerfelli, 1998],
- regrouper ces paragraphes en unités documentaires sémantiques [Ouerfelli, 2000],
- déterminer la structure logique.

La phase d'extraction se termine par l'analyse lexicale des étiquettes des arborescences précédemment définies. L'analyse lexicale consiste à résoudre les problèmes de synonymie pour des documents ayant des arborescences similaires [Miller, 1995] et d'unicité de noms au sein d'une même arborescence.

3.5 PHASE DE COMPARAISON

Cette phase consiste à comparer la structure arborescente d'un document source à celles de l'entrepôt afin de déterminer s'il existe une structure logique générique identique ou approchante [Khrouf et al., 2003]. Si une structure identique ou approchante existe, le système *fusionne* les deux structures en une structure logique générique et rattache le document à cette structure. Dans le cas contraire, le système *crée* une nouvelle structure logique générique afin d'y rattacher le document. L'approche proposée lors de cette phase est basée sur le calcul de similarité d'arborescences hétérogènes d'éléments ordonnés et étiquetés. Il s'agit de spécifier les opérations de fusion de schémas (structures logiques) pouvant être réalisées. **La spécificité de ces travaux provient du fait que la structure logique du document est une organisation hiérarchique d'éléments étiquetés et ordonnés.**

3.5.1 Les étapes de la phase de comparaison

Pour répondre aux besoins de cette phase de comparaison, nous avons défini six étapes schématisées comme suit [Khrouf et al., 2003] :

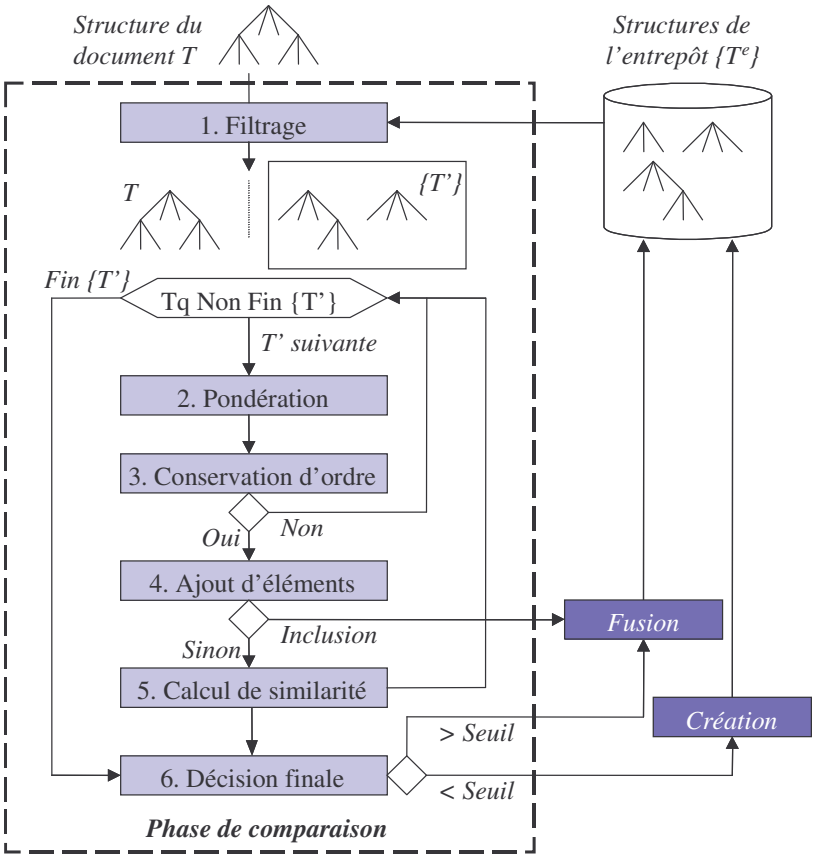


Figure 14 : Phase de comparaison

Ces étapes sont définies comme suit :

1. **Filtrage** : cette étape consiste à sélectionner l'ensemble des structures logiques génériques de l'entrepôt pouvant être fusionnées (T') avec la structure logique du document à intégrer (T). Cette étape calcule un rapport de filtrage évaluant le nombre de nœuds communs entre deux structures comparées. Les structures génériques pour lesquelles ce rapport est supérieur à un seuil (déterminé par expérimentation), sont sélectionnées pour les étapes suivantes. Pour chaque structure logique générique sélectionnée (dans l'ordre décroissant du nombre d'éléments communs), le système doit effectuer les étapes suivantes (2 à 5).

2. **Pondération** : cette étape affecte des poids aux différents nœuds des deux arborescences (structure logique du document *T* et structure logique générique sélectionnée dans l'entrepôt *T'*) en fonction de leur position dans l'arborescence. Cette mesure de pondération tient compte de la profondeur (niveau d'arborescence) et de la largeur de l'arborescence (ordre des fils d'un élément père).

3. **Conservation d'ordre** : cette étape permet d'étudier le placement des nœuds communs aux deux structures à comparer. Pour chacun des nœuds communs, cette étape vérifie (1) qu'ils possèdent les mêmes ancêtres communs (2) qu'ils possèdent les mêmes fils communs et que l'ordre de ces fils est identique dans les deux arborescences. Si l'une des deux règles n'est pas vérifiée, les deux structures arborescentes sont considérées suffisamment différentes pour générer l'arrêt du processus de comparaison et étudier la structure logique générique suivante.

4. **Ajout d'éléments** : cette étape consiste à ajouter des éléments dans l'une des deux structures arborescentes pour les homogénéiser. Pour chaque nœud commun à T ou T', cette étape permet d'ajouter des nœuds ancêtres ou fils. Si les deux structures sont

devenues identiques, le document est rattaché à la structure logique générique de l'entrepôt (arrêt de la phase de comparaison). Dans le cas contraire, l'étape suivante est déclenchée.

5. **Calcul de similarité** : cette étape consiste à calculer un degré de similarité entre la structure logique du document (T) et la structure logique générique sélectionnée dans l'entrepôt lors du filtrage (T'). Le calcul de similarité entre T et T' repose sur la distance d'alignement des différents nœuds communs aux deux structures arborescentes. La distance d'alignement entre deux nœuds communs, correspond à la valeur absolue de la différence des poids de ces nœuds.

6. **Décision finale** : cette étape prend la décision de fusionner ou non la structure du document avec une des structures logiques génériques de l'entrepôt en se basant sur les degrés de similarité calculés préalablement. Cette étape sélectionne la structure logique générique de l'entrepôt dont le degré de similarité est le plus élevé puis le compare à un seuil de fusion (déterminé par expérimentation). Si le degré de similarité est strictement inférieur à ce seuil, une nouvelle structure logique générique sera créée à partir de la structure logique extraite du document à intégrer. Dans le cas contraire, la fusion entre T et T' est possible : les éléments occupant la même position dans les deux arborescences seront remplacés par un seul élément complexe. Un élément complexe est une composition d'éléments simples interconnectés par l'opérateur "ou".

Tous les détails de cette phase sont explicités dans [Khrouf et al., 2003] et [Khrouf, 2004] avec notamment les définitions précise des différents indicateurs et seuils.

3.5.2 Exemple d'application

Nous avons un document source avec une structure T que nous devons comparer à une structure générique de l'entrepôt T'.

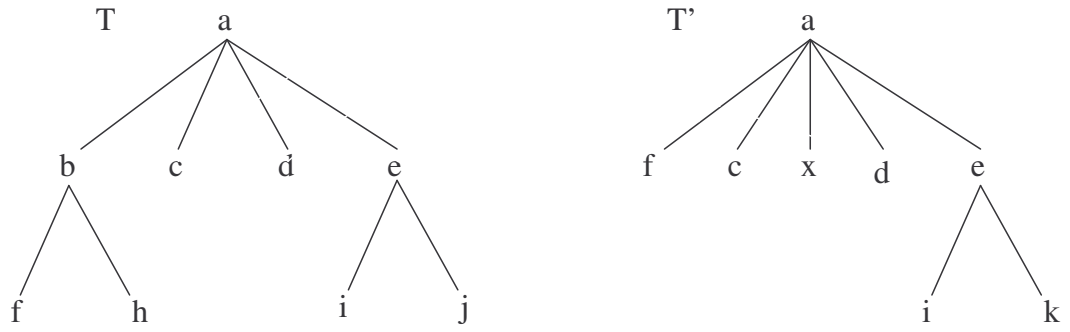


Figure 15 : Structures à comparer

L'étape de filtrage indique que ces deux arborescences ont 71% de nœuds en commun. Sachant que suite à nos expérimentations, le seuil de filtrage était fixé à 60%, nous avons pu effectuer les étapes de pondération et de conservation d'ordre. Les deux règles de conservation d'ordre des fils et des ancêtres étant respectées, le processus de comparaison s'est poursuivi par l'ajout d'éléments. Cette étape a modifié T par l'ajout un nœud fils "x" au nœud "a". La modification de T' s'est traduite par l'ajout d'un nœud intermédiaire "b" entre "a" et "f" et l'ajout d'un fils "h" à "b". Les deux structures étant encore différentes (j et k non alignés), l'étape du calcul de similarité a abouti à un rapport de 0,93. Ce degré de similarité est supérieur au seuil (0,75 déterminé lors de nos expérimentations), le système a proposé de fusionner les deux structures précédemment étudiées. Cette fusion consiste à modifier la structure de T' de manière à intégrer un nœud complexe intégrant une alternative entre les étiquettes j et k.

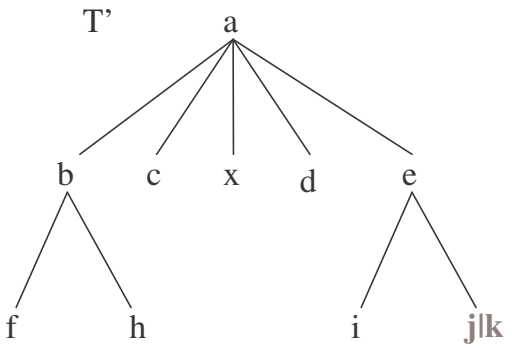


Figure 16 : Arborescence finale de T'

4 BILAN ET PERSPECTIVES

Un entrepôt constitue la zone de préparation des données décisionnelles. La problématique majeure des entrepôts est de centraliser et d'historiser les données sources hétérogènes de manière à faciliter les processus d'analyses décisionnelles. Dans ce domaine, les premiers travaux se sont centrés sur des aspects techniques tels que l'alimentation d'entrepôts relationnels au travers des techniques de vues matérialisées et de médiateurs. Or, au début de nos travaux de recherches, il n'y avait pas de solution pour aider un concepteur d'entrepôts. Les travaux exposés dans ce chapitre apportent une **solution pour la conception d'entrepôts** intégrant des données factuelles ou des données documentaires.

4.1 Bilan sur la modélisation des entrepôts

Dans le cadre des ED, de nombreux auteurs [Inmon, 1996 ; Chaudhuri & Dayal, 199 ; Yang & Widom, 1998 ; Pedersen & Jensen, 1999 ; Yang & Widom, 2000] s'accordent sur le fait que la gestion de l'évolution des données au cours du temps est vitale pour les décideurs. Or, à notre connaissance, il n'y a pas eu de proposition précise pour un modèle conceptuel d'entrepôts de données évolutives.

Pour répondre à ce besoin, nous avons étendu le modèle objet standard afin d'intégrer les concepts de filtre temporel, de filtre d'archivage, de fonction de dérivation et d'environnement. Un **filtre temporel** précise les attributs des **objets entrepôt** pour lesquels l'évolution des valeurs est conservée à chaque extraction de données sources. Le **filtre d'archivage** permet d'agréger l'ensemble des valeurs temporisées d'un même attribut quand la conservation des valeurs de manière détaillée n'est plus utile pour la prise de décision. La **fonction de dérivation** permet d'exprimer au travers de primitives algébriques le processus d'extraction des données sources permettant d'alimenter les données de l'entrepôt. Enfin, le concept d'**environnement** permet de définir différents espaces évolutifs dans un même ED et précise le niveau de granularité d'historisation. Ces propositions conceptuelles ont été complétées par le développement d'un outil d'aide à la conception d'ED (cf. chapitre VI). Cet outil permet de construire graphiquement un schéma conceptuel et traduit chaque classe en un ensemble de tables relationnelles.

De plus, les ED construits uniquement à partir de BD sources n'extraient que 20% des informations disponibles et les 80% restants sont stockés dans des documents [Tseng & Chou, 2006]. Aussi, dans un second temps, nous avons apporté des solutions pour la conception d'entrepôts de documents (EDO). Dans ce cadre, la problématique à traiter est l'intégration de documents hétérogènes en structure et en contenu afin de pouvoir effectuer des traitements décisionnels. Afin de **supporter la nature hétérogène des documents sources**, l'intérêt majeur de nos travaux est que nous n'imposons pas la définition d'une structure documentaire spécifique au préalable. Notre approche repose sur la **classification des documents** textuels en fonction de leur structure logique. Nos solutions reposent sur deux principes majeurs :

- une modélisation **générique d'entrepôts de documents textuels** permettant de représenter les structures logiques des documents ainsi que leurs contenus. Ce modèle générique présente l'avantage de pouvoir être instancié à partir de documents de nature hétérogène tels que des documents structurés (XML valide), semi-structurés (HTML) ou non structurés (TXT) ;
- unprocessus d'alimentation basé sur l'extraction de la structure logique spécifique au document à intégrer ainsi que de son contenu. Ce processus d'alimentation est basé sur un **algorithme de comparaison d'arbres étiquetés et ordonnés** représentant une structure logique.

4.2 PRODUCTION SCIENTIFIQUE

Le travail présenté dans ce chapitre a donné lieu à plusieurs encadrements et co-encadrements. Notamment, les travaux relatifs à la gestion optimale des données évolutives d'un ED se sont concrétisés par le co-encadrement (à 80%) de la thèse d'Olivier Teste [Teste, 2000] et l'encadrement de 3 travaux de recherches du DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier. Le premier, Xavier Baril [Baril, 1999], a permis de valider le principe de l'historisation. Le second, Faiza Ghozzi [Ghozzi, 2000], a permis de valider les filtres temporels et d'archivages. Le troisième, Tarek Jarraya [Jarraya, 2001] s'est concentré sur les principes de dérivation des données sources.

Nos travaux sur l'intégration de documents dans des entrepôts se sont déroulés dans le cadre de collaborations avec d'autres membres de l'IRIT : Chantal Soulé-Dupuy et Kais Khrouf de la composante D2S2 (Documents, Données Semi-Structurées et usages) de l'équipe SIG ainsi que Josiane Mothe et Farshad Riahi de la composante EVI (Exploration et Visualisation d'Information) de l'équipe SIG. Cette collaboration s'est concrétisée par le co-encadrement (50%) thèse de Kais Khrouf [Khrouf, 2004]. Actuellement, ces travaux ont été étendus par les chercheurs de la composante D2S2 afin d'intégrer des documents multimédia.

Nos travaux ont été également supportés par deux projets. Le projet REANIMATIC s'est effectué en collaboration avec l'association de médecins OUTCOME-REA et plusieurs équipes de recherche (PRISM, IRIT, LIRMM…). Ce projet a consisté à développer une modélisation spécifique des entrepôts de données médicales évolutives (collectées à partir des bases opérationnelles des services de réanimation) afin d'améliorer la qualité des soins et le devenir des patients dans les services de réanimation des hôpitaux français. Ce projet a servi de cadre applicatif aux travaux de thèse d'Olivier Teste (soutenue en 2000). De plus, une convention de recherche entre l'Université Paul Sabatier (équipe SIG) et le CTI-Sud (Centre de Traitement Informatique de l'Assurance Maladie à Toulouse) a été conclue en 1998. Cette étude a eu pour but de spécifier un entrepôt de données médicales alimenté par la base SIAM (Système d'Information de l'Assurance Maladie).

D'un point de vue publications, nous pouvons citer les références suivantes[10] :

- 1 article dans l'ouvrage international "Entreprise Information Systems II" (sélection des meilleurs articles de ICEIS'00) [Ravat &Teste, 2000a],
- 4 articles dans des conférences internationales : CIKM'99 [Ravat et al. 1999], ECIS'00 [Mothe et al., 2000], DEXA'00 [Ravat & Teste, 2000b] et ADBIS'00 [Ravat & Teste, 2000c],
- 1 article dans la revue nationale RSTI-ISI [Khrouf et al., 2003],

[10] Le contenu de chaque article est résumé dans le chapitre 6.

- 2 articles dans les conférences nationales BDA'00 [Ravat et al., 2000] et BDA'01 [Ravat & Teste, 2001]

4.3 PERSPECTIVES

Du fait de l'encadrement de deux thésards issus de composantes différentes de l'IRIT, les travaux que nous avons menés dans le domaine des ED et des EDO ont été effectués de manière séparée.

Une première extension à ces travaux serait d'intégrer les concepts définis dans ces deux travaux afin de pouvoir profiter des avantages de chacun d'eux (gestion de données évolutives et intégration de documents hétérogènes). Une première perspective consiste à intégrer l'évolution des documents en travaillant avec plusieurs versions. Nous avons apporté une première solution dans [Khrouf et al., 2007b]. Une deuxième perspective consiste à proposer des mécanismes d'archivage adaptés aux documents comme la génération automatique de résumés [Radev et al., 2002] ou la catégorisation automatique de documents en fonction de hiérarchie de mots-clés [Chakrabarti et al., 1998 ; Agrawal et al., 2000].

Une deuxième extension pourrait permettre de gérer au mieux l'important volume des données des entrepôts. Pour ce faire, nous pourrions profiter de mon expérience dans la conception de BD réparties acquises durant ma thèse [Ravat, 1996] pour proposer des mécanismes adaptés pour la répartition des données d'un entrepôt. Cette répartition doit tenir compte des besoins des différents magasins, de la localisation et la disponibilité des sources et des temps d'accès aux données. Il faut donc proposer de nouveaux algorithmes de fragmentation de données décisionnelles et de nouvelles solutions pour l'allocation de données de base ou agrégées.

CHAPITRE III :

MODELISATION DES

MAGASINS DE DONNES

PLAN DU CHAPITRE

1 INTRODUCTION A LA MODELISATION DES MAGASINS DE DONNEES 53

2 MODELE GENERIQUE DE BASE 54
2.1 Problématique 54
2.2 Dimension 55
2.3 Hiérarchie 55
2.4 Fait 56
2.5 Constellation 57
2.6 Table multidimensionnelle 58

3 INTEGRATION DE DONNEES TEXTUELLES 59
3.1 Problématique 59
3.2 Typologie des mesures 60
3.3 Dimensions représentant un document 60

4 GESTION DE LA COHERENCE SEMANTIQUE 62
4.1 Problématique 62
4.2 Typologie des contraintes 62
4.3 Contraintes sémantiques intra-dimensions 63
4.4 Contraintes sémantiques inter-dimensions 64

5 GESTION DE LA COHERENCE TEMPORELLE 66
5.1 Problématique 66
5.2 Principes 67
5.3 Constellation et versions d'étoile 68
5.4 Composants d'une version d'étoile 69

6 INTEGRATION ET CAPITALISATION DE L'EXPERTISE DES DECIDEURS 70
6.1 Problématique 71
6.2 Principes 71
6.3 Les annotations décisionnelles 72
6.4 Ancrage d'annotations décisonnelles 73
6.4.1 Annotation d'un schéma multidimensionnel 73
6.4.2 Annotation d'une Table Multidimensionnelle (TM) 74

7 PERSONNALISATION DE MAGASINS DE DONNEES 74
7.1 Problématique 75
7.2 Nos résultats 75
7.2.1 Approche "naïve" 76
7.2.2 Approche "avancée" 76

8 BILAN ET PERSPECTIVES 78
8.1 Bilan sur la modélisation des magasins 79
8.2 Production scientifique 79
8.3 Perspectives 79

1 INTRODUCTION A LA MODELISATION DES MAGASINS DE DONNEES

Dans ce chapitre, notre objectif est de présenter les travaux que nous avons réalisés dans le cadre des magasins de données reposant sur une modélisation multidimensionnelle des données. Ce modèle permet de se centrer sur les besoins décisionnels en représentant les données par sujets analysés selon différentes dimensions.

Comme indiqué dans [Rizzi et al., 2006] et [Niemi, et al. 2003], la modélisation OLAP des données ne repose pas sur une formalisation précise, stable et reconnue par l'ensemble de la communauté scientifique. L'ensemble des propositions reste parcellaire. L'objectif de nos travaux menés dans le cadre de la modélisation multidimensionnelle vise à combler ce manque. Plus précisément, notre objectif est d'offrir une solution pour la **modélisation multidimensionnelle orientée décideur**. Notamment, nous proposons de décliner cette orientation utilisateur final en plusieurs sous-objectifs :

- Dans un premier temps, nous souhaitons proposer un modèle **conceptuel générique orienté décideur**. L'aspect conceptuel permettra de se concentrer sur les éléments à analyser en faisant abstraction de leur implantation. L'aspect générique permettra de pouvoir représenter au sein d'un même modèle l'ensemble des concepts proposés jusqu'à maintenant dans diverses propositions. De même, l'aspect générique doit permettre de représenter de manière homogène les indicateurs numériques et les indicateurs textuels issus des sources documentaires. L'orientation décideur doit se traduire par la spécification précise aussi bien des structures de représentation multidimensionnelle que les structures de visualisation des analyses décisionnelles.

- Les décideurs souhaitent un modèle leur permettant de représenter l'ensemble des concepts multidimensionnels mais ils souhaitent que les schémas qu'ils manipulent reposent sur une modélisation fiable des données. Pour répondre à ce besoin, nous souhaitons proposer une solution pour une **modélisation de la cohérence sémantiquement et temporellement** des données décisionnelles. La fiabilité sémantique permettra à un décideur de manipuler des données correctes et de pouvoir effectuer des croisements de données ayant un sens. La fiabilité temporelle permettra à un décideur de pouvoir effectuer des analyses sur des schémas multidimensionnels évoluant dans le temps en fonction des changements des sources de données ou des besoins des décideurs.

- Seules, les données brutes (même fiables sémantiquement et temporellement) ne sont pas suffisantes pour prendre des décisions au sein d'une organisation. A l'heure actuelle, il n'y a pas sur le marché de modèle et d'outil informatique permettant de représenter dans un même espace de stockage aussi bien les données multidimensionnelles (patrimoine matériel) que la **matérialisation du savoir-faire du décideur** (patrimoine immatériel). Les décideurs souhaitent également accompagner ces données de commentaires ou de remarques et souhaitent également faire partager ces données complémentaires à d'autres décideurs afin de prendre en compte leurs avis avant de prendre une décision. Pour répondre à ce besoin, notre objectif est d'offrir une **Mémoire d'Expertise Décisionnelle (MED)**. Une telle mémoire sauvegarde aussi bien des données multidimensionnelles fiables que des commentaires, remarques pouvant expliciter une analyse décisionnelle.

- De plus, afin de faciliter la tâche du décideur, nous souhaitons proposer des mécanismes de **personnalisation** des schémas multidimensionnels. Cette personnalisation permettra de mettre en avant les données considérées comme les plus significatives par les décideurs et facilitera les analyses OLAP.

Dans les sections suivantes, nous présentons les résultats que nous avons obtenus en réponse à ces différentes problématiques. La section 2 présente les concepts du modèle multidimensionnel générique que nous proposons. La section 3 met en avant l'intégration de données textuelles dans un modèle multidimensionnel. Les sections 4 et 5 permettent de se centrer sur la gestion de la cohérence sémantique et temporelle. Les sections 6 et 7 étudient l'intégration du savoir-faire décisionnel dans un schéma OLAP ainsi que sa personnalisation.

2 MODELE GENERIQUE DE BASE

En l'absence d'un modèle multidimensionnel reconnu et stable, les propositions actuelles offrent des solutions à différents niveaux d'abstraction (conceptuel, logique et physique) et ne supportent que partiellement l'ensemble des concepts multidimensionnels présentés dans le premier chapitre. Les travaux de recherche exposés dans cette section visent à combler dans un premier temps ces lacunes. La section suivante présente plus en détail la problématique et les autres sections présentent nos résultats.

2.1 PROBLEMATIQUE

En l'état actuel, nous pouvons classer les propositions en deux catégories [Pedersen et al., 2001], [Torlone, 2003] et [Abelló et al., 2006]. La première intitulée « Modèle Cube » [Agrawal et al., 1995 ; Li & Wang, 1996 ; Gyssens et al., 1996 ; Agrawal et al., 1997 ; Gyssens & Lakshmanan, 1997 ; Datta & Thomas, 1999 ; Lehner, 1998] propose de représenter les données sous forme de cube sans expliciter les différents composants d'un schéma multidimensionnel. La seconde catégorie, appelée "Modèle multidimensionnel", est sémantiquement plus riche car elle permet d'expliciter précisément les différents composants d'un schéma multidimensionnel [Pedersen et al., 2001 ; Abelló et al., 2003 ; Abelló et al., 200 ; Cabibbo & Torlone, 1998 ; Cabibbo & Torlone, 2000 ; Torlone, 2003 ; Schneider, 2007].

En réponse à notre objectif, nous nous plaçons dans la seconde catégorie. Cependant, la plupart des modèles existants proposent de définir des schémas en étoile basés sur la dualité fait-dimensions [Kimball & Ross, 2002]. Cette modélisation en étoile ne facilite pas la corrélation entre les différents sujets d'analyse tels que, par exemple, la comparaison entre les ventes et les achats annuels d'un même produit dans une application d'analyse commerciale. De plus, tous les travaux n'intègrent pas des dimensions avec une définition explicite de différentes perspectives d'analyses (hiérarchies). Or, la multiplicité des hiérarchies est une caractéristique du monde réel ; souvent les objets sont classés selon différents critères indépendants (tranches d'âge et adresse pour les clients par exemple).

Afin de pallier ces limites, nos travaux ont permis de proposer un modèle conceptuel multidimensionnel générique. Cette solution présente les deux avantages suivants :

- **le niveau conceptuel** permet d'avoir une vision proche des décideurs tout en faisant abstraction de toute implantation physique particulière et de toute modélisation logique facilement déductible à partir d'une représentation conceptuelle ;
- **le modèle multidimensionnel générique** permet de supporter une définition explicite de l'ensemble des concepts inhérents à la représentation et à la manipulation OLAP des données. Pour la représentation, nous proposons une modélisation multi-sujets analysés à l'aide d'axes intégrant des vues d'analyses multiples. Pour la manipulation, nous spécifions un concept permettant une visualisation facilement exploitable par les décideurs.

Dans les sections suivantes, nous définissons les différents concepts et formalismes de notre modèle multidimensionnel générique.

2.2 DIMENSION

Une dimension modélise un axe d'analyse en fonction duquel peuvent être observées les valeurs analysées. Une dimension est caractérisée par des attributs ; chaque attribut représente une façon de graduer l'axe d'analyse. Les différents attributs d'une dimension sont organisés au sein d'une ou plusieurs vues appelées hiérarchies.

Définition. Une dimension D_i est définie par (N^{Di}, A^{Di}, H^{Di}, I^{Di}) où

- N^{Di} est le nom de la dimension,
- $A^{Di} = \{a^{Di}_1, a^{Di}_2, \ldots, a^{Di}_u\}$ est un ensemble d'attributs,
- $H^{Di} = \{h^{Di}_1, h^{Di}_2, \ldots, h^{Di}_y\}$ est un ensemble de hiérarchies,
- $I^{Di} = \{I^{Di}_1, I^{Di}_2, \ldots\}$ est l'ensemble des instances de D_i.

Exemple. Une entreprise de location de véhicules souhaite analyser les résultats de ces différentes agences réparties en France et aux Etats unis. L'analyse des locations doit tenir compte du jour de la location, du client, de l'agence et du véhicule. Pour répondre à ce besoin, nous proposons de spécifier quatre dimensions (TEMPS, CLIENTS, VOYAGES et AGENCES). Chaque agence est caractérisée par son code, sa raison sociale et sa localisation. La dimension AGENCES est définie comme suit :

- $N^{AGENCES}$ = "AGENCES",
- $A^{AGENCES}$ = {CodeAg, Raison, Ville, Département, Nom_dpt, Région, Pays, Etat, Zone},
- $H^{AGENCES} = \{ h^{AGENCES}_1, h^{AGENCES}_2, h^{AGENCES}_3, h^{AGENCES}_4 \}$,
- $I^{AGENCES} = \{ I^{AGENCES}_1, I^{AGENCES}_2, I^{AGENCES}_3, \ldots \}$.

Chaque instance est un n-uplet de la forme suivante :

- $I^{AGENCES}_1$ = [CodeAg : 1, Raison : "Agence Campus31", Ville : "Toulouse", Département : 31, Nom_dpt : "Hte-Garonne", Région : "Midi-Pyrénées", Pays : "France", Etat : NULL, Zone : 'Sud-FR'],
- $I^{AGENCES}_2$ = [CodeAg : 2, Raison : "Agence du Bouchon", Ville : "Lyon", Département : 69, Nom_dpt : "Rhône", Région : "Rhône-Alpes", Pays : "France", Etat : NULL, Zone : 'Est-FR'],
- $I^{AGENCES}_3$ = [CodeAg : 3, Raison : "Big Appel Agency", Ville : "New York", Département : NULL, Nom_dpt : NULL, Région : NULL, Pays : "Etats-Unis", Etat : "New York", Zone : 'Est-EU'].

2.3 HIERARCHIE

Une hiérarchie représente une perspective d'analyse précisant les niveaux de granularité en fonction desquels peuvent être observés les indicateurs d'analyse. Une hiérarchie h^{Di}_x définie sur la dimension D_i est un chemin élémentaire acyclique débutant par l'attribut de plus forte granularité et se terminant par un attribut de plus faible granularité (qui est le plus souvent un identifiant).

Définition. Une hiérarchie h^{Di}_x est définie par (N^{Di}_x, $Param^{Di}_x$, $Suppl^{Di}_x$, $Cond^h$) où

- N^{Di}_x est le nom de la hiérarchie,
- $Param^{Di}_x = <a^{Di}_{k0}, a^{Di}_{k1},\ldots, a^{Di}_{kz}>$ est un ensemble ordonné décrivant la hiérarchie des attributs de la plus forte granularité vers la plus faible (chaque attribut est appelé paramètre de la hiérarchie et correspond à un niveau de granularité d'analyse),
- $Suppl^{Di}x$: $Param^{Di}_x \rightarrow 2^{ADi - ParamDix}$ est une application spécifiant les attributs faibles qui complètent la sémantique des paramètres (chaque paramètre est associé à un ensemble d'attributs faibles),
- $Cond^h$ est une expression booléenne définissant la condition d'appartenance des instances de la dimension à une hiérarchie.

Exemple. Nous souhaitons compléter la définition de la dimension AGENCES en lui associant trois perspectives d'analyse en fonction de la localisation de l'agence. La première perspective décrit les agences suivant l'organisation géographique de la France en ville, département et région. La deuxième perspective, relative à l'organisation géographique des Etats-Unis, organise les villes par état. Enfin, la troisième perspective, commune à la France et aux Etats-Unis, décrit la position géographique des villes dans leur pays selon l'indication nord, sud, est, ouest. Ces différentes perspectives d'analyse sont représentées à l'aide de trois hiérarchies ($h^{AGENCES}_1$, $h^{AGENCES}_2$, et $h^{AGENCES}_3$) définies comme suit :

- $h^{AGENCES}_1$ = ("geo_fr", {$Param^{AGENCES}_1$(CodeAg) = Ville, $Param^{AGENCES}_1$(Ville) = Département, $Param^{AGENCES}_1$ (Département) =Région, $Param^{AGENCES}_1$ (Région) =Pays}, {$Suppl^{AGENCES}_1$(CodeAg) = {Raison}, $Suppl^{AGENCES}_1$(Département) = {Nom_dpt}}, Pays = "France"),
- $h^{AGENCES}_2$ = ("geo_us", {$Param^{AGENCES}_2$(CodeAg) = Ville, $Param^{AGENCES}_2$(Ville) = Etat, $Param^{AGENCES}_2$(Etat) = Pays}, {$Suppl^{AGENCES}_2$(CodeAg) = {Raison}}, Pays = "Etats-Unis" $\wedge$ Etat $\neq$ NULL),
- $h^{AGENCES}_3$ = ("geo_zn", {$Param^{AGENCES}_3$(CodeAg) = Ville, $Param^{AGENCES}_3$(Ville) = Zone, $Param^{AGENCES}_3$(Zone) = Pays}, {$Suppl^{AGENCES}_3$(CodeAg) = {Raison}}, Zone $\neq$ NULL),

La spécificité de multi-instanciation de notre modèle réside dans l'intégration d'une condition d'appartenance des instances de la dimension aux hiérarchies. Ainsi,

- les instances {$I^{AGENCES}_1$, $I^{AGENCES}_2$} appartiennent à $h^{AGENCES}_1$ tandis que
- l'instance {$I^{AGENCES}_3$} appartient à $h^{AGENCES}_2$ et
- les instances {$I^{AGENCES}_1$, $I^{AGENCES}_2$, $I^{AGENCES}_3$} appartiennent à $h^{AGENCES}_3$.

2.4 FAIT

Un fait modélise un sujet d'analyse composé d'un ensemble d'indicateurs appelés mesures. Les mesures sont le plus souvent numériques et additives (ou semi-additives) (Kimball, 1996).

Définition. Un fait F_j est défini par (N^{Fj}, M^{Fj}, I^{Fj}, $IStar^{Fj}$) où

- N^{Fj} est le nom du fait,
- M^{Fj} = { $f_1(m_1)$, … f_w (m_w)} est un ensemble de mesures avec les fonctions d'agrégation associées,
- I^{Fj} = {I^{Fj}_1, I^{Fj}_2,…} est l'ensemble des instances de F,
- $IStar^{Fj}$ est une fonction associant chaque instance de I^{Fj} à une instance de chaque dimension liée au fait.

Exemple. Nous complétons l'exemple précédent afin de représenter le fait LOCATION. Ce fait doit permettre d'analyser le montant et le nombre de jours de chaque location. La définition ci-dessous permet de représenter ce fait avec ces deux mesures et deux instances.

- $N^{LOCATION}$ = "LOCATION",
- $M^{LOCATION}$ = {montant, nbjours},
- $I^{LOCATION}$ = $\{I^{LOCATION}_1, I^{LOCATION}_2,\ldots\}$,
- $IStar^{LOCATION}$ est définie par $\{I^{LOCATION}_1 \rightarrow \{I^{TEMPS}_1, I^{VEHICULE}_1, I^{AGENCES}_1, I^{CLIENTS}_1\}$, $I^{LOCATION}_2 \rightarrow \{I^{TEMPS}_1 I^{VEHICULE}_2, I^{AGENCES}_1, I^{CLIENTS}_1\},\ldots\}$.

Notons que les 2 instances du fait LOCATION ($I^{LOCATION}_1$, $I^{LOCATION}_2$) concernent un client ($I^{CLIENTS}_1$) se rendant dans une agence ($I^{AGENCES}_1$) pour louer le même jour (I^{TEMPS}_1) deux véhicules ($I^{VEHICULE}_1$ et $I^{VEHICULE}_2$). Chaque instance de fait est un n-uplet de la forme suivante :

- $I^{LOCATION}_1$ = [montant : 540.00, nbjours : 8],
- $I^{LOCATION}_2$ = [montant : 1200.00, nbjours : 22].

2.5 CONSTELLATION

Une constellation est la généralisation du modèle en étoile (Kimball, 1996) dans lesquels un seul fait (sujet d'analyse) est modélisé. Une constellation permet de représenter plusieurs faits associés à des dimensions éventuellement partagées.

Définition. Une constellation CS est définie par (N^{CS}, F^{CS}, D^{CS}, $Star^{CS}$) où

- N^{CS} est le nom de la constellation,
- $F^{CS} = \{F_1, F_2,\ldots, F_p\}$ est un ensemble de faits,
- $D^{CS} = \{D_1, D_2,\ldots, D_q\}$ est un ensemble de dimensions,
- $Star^{CS} : F^{CS} \rightarrow 2^{DCS}$ est une fonction associant les faits aux dimensions afin de spécifier les sujets d'analyses et les axes d'étude associés.

Exemple. Sachant que la direction souhaite corréler l'analyse des locations définie précédemment avec l'étude les performances (chiffre d'affaire et marge) des employés de chaque agence, nous proposons d'étendre le schéma multidimensionnel précédent. Pour répondre à ce besoin, nous proposons de spécifier une constellation comprenant deux faits (LOCATION, PERF) et cinq dimensions (TEMPS, CLIENTS, AGENCES, EMPLOYES, VEHICULES). La constellation est définie comme suit :

- N^{CS} = "LOUEVOYAGE",
- F^{CS} = {LOCATION, PERF},
- D^{CS} = {TEMPS, EMPLOYES, VEHICULES, AGENCES, CLIENTS},
- $Star^{CS}$ = {VENTES →{TEMPS, VEHICULES, AGENCES, CLIENTS},
  PERF →{ TEMPS, EMPLOYES, AGENCES} },

Afin d'être plus expressif, nous avons associé à ce formalisme textuel un formalisme graphique permettant de mettre en évidence les différents composants d'un schéma multidimensionnel. Ce formalisme graphique est basé sur une adaptation du formalisme introduit par [Golfarelli et al., 1998].

Exemple. la figure suivante illustre la constellation définie précédemment

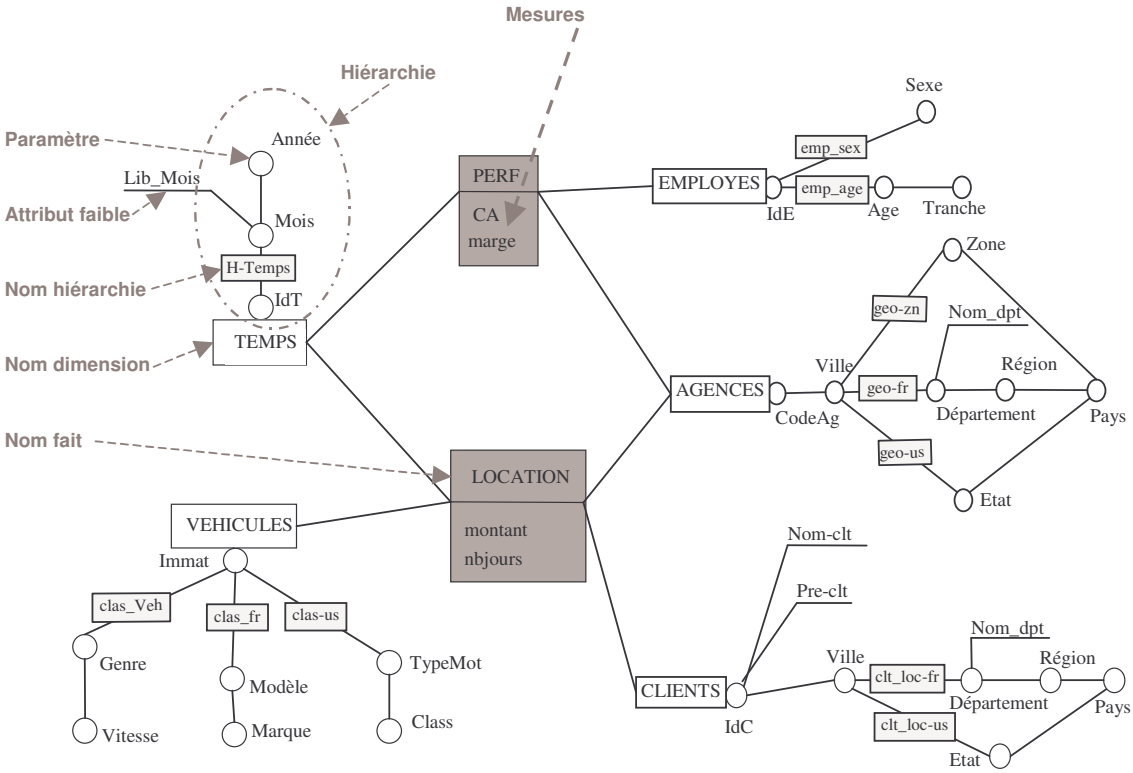


Figure 17 : Exemple d'une représentation graphique d'une constellation

2.6 TABLE MULTIDIMENSIONNELLE

Comme indiqué dans le premier chapitre, les nombreux travaux relatifs à la manipulation OLAP ne définissent pas toujours une structure de visualisation des données décisionnelles. Sachant qu'une visualisation sous forme de cubes en N dimensions (N>2) semble difficilement exploitable par les décideurs [Gyssens & Lakshmanan, 1997] et que ce type de présentation fait abstraction de la hiérarchisation des dimensions, nous avons proposé le concept de Table Multidimensionnelle (TM). Cette structure de visualisation consiste à représenter les données analysées sous forme de tableau à double entrées hiérarchisées. Cette visualisation centre l'analyse sur un seul fait. Ce type de restitution réduit la complexité de l'information visualisée et facilite l'interprétation et l'analyse des données ; il s'agit d'une représentation très répandue [Gyssens & Lakshmanan, 1997] dans les différents outils décisionnels.

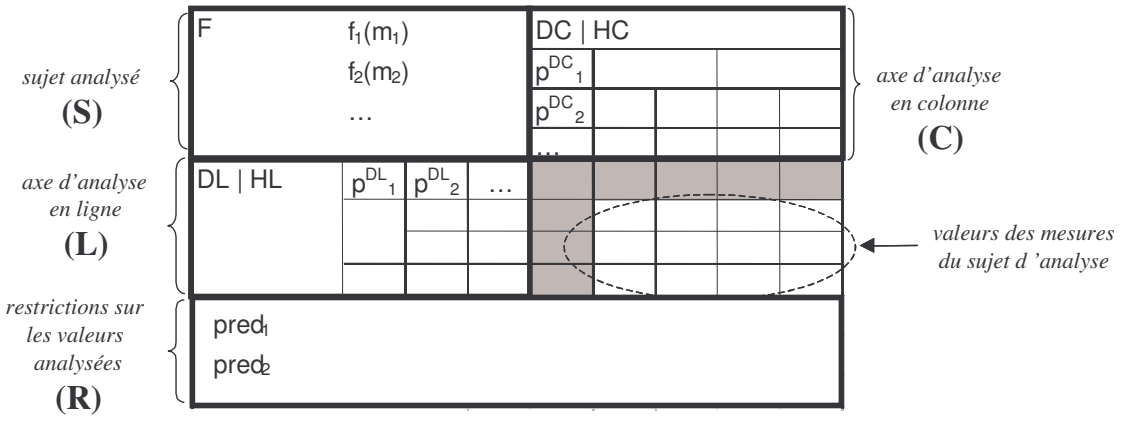


Figure 18 : Structure d'une table multidimensionnelle

Définition. Une table multidimensionnelle T est définie par (S, L, C, R) où

- S = (F, {$f_1(m_1)$, … $f_p(m_u)$}) représente le sujet analysé dans T au sein d'un fait courant F au travers duquel sont observées des mesures {m_1, … m_u} agrégées à l'aide de fonctions f_1, … f_u,
- L = (DL, HL, <p^{DL}_1, …, p^{DL}_v>) représente l'axe d'analyse en ligne de la table T au travers d'une dimension courante DL, d'une hiérarchie courante HL et d'une liste ordonnée de paramètres affichés <p^{DL}_1, p^{DL}_2,…>,
- C = (DC, HC, <p^{DC}_1,…, p^{DC}_w>) représente l'axe d'analyse en colonne de la table T au travers d'une dimension courante DC, d'une hiérarchie courante HC et d'une liste ordonnée de paramètres affichés <p^{DC}_1, p^{DC}_2,…>,
- R = $pred_1 \cup \ldots \cup pred_x$ est une conjonction de prédicats normalisés portant sur les dimensions courantes, les autres dimensions liées au fait F et éventuellement sur le fait lui-même.

Exemple. Si un décideur souhaite analyser le montant des ventes des voyages pour les trois dernières années en fonction du pays d'origine des clients, il doit construire la TM suivante :

VENTES Sum(montant)		CLIENTS | Cli_geo		
		pays	Etats Unis	France
TEMPS | H_an	Annee			
	2007		***200***	***150***
	2006		***250***	***240***
	2005		***200***	***210***

Figure 19 : Exemple d'une table multidimensionnelle

3 INTEGRATION DE DONNEES TEXTUELLES

Un des points forts de nos travaux a été de pouvoir intégrer dans des entrepôts aussi bien des données sources factuelles que documentaires. Cette spécificité a un impact sur la modélisation multidimensionnelle des magasins de données. Les questions qu'il faut se poser sont : est-il possible d'analyser des données documentaires à l'aide d'un schéma multidimensionnel? Faut-il apporter des modifications voire des extensions au modèle générique défini précédemment?

Dans les sections suivantes, nous précisons la problématique liée à l'intégration de données documentaires et nous exposons les résultats que nous avons obtenus.

3.1 PROBLEMATIQUE

L'analyse de documents textuels permet à un utilisateur d'obtenir une vision globale de l'ensemble d'une collection de documents [McCabe et al., 2000]. Notre objectif est de pouvoir représenter de manière multidimensionnelle les données relatives aux documents. Notamment, notre souhait est de pouvoir représenter les différentes informations extraites des documents (contenu, structure logique et méta-informations) [Ravat et al., 2007e]. L'intégration de ces différentes données aura un impact sur les composants d'un schéma multidimensionnel.

Pour répondre à ce besoin, nous avons proposé un modèle multidimensionnel étendu pour l'analyse multidimensionnelle de données textuelles. Cette extension intervient à deux niveaux :

- Spécification du type des mesures,
- Définition de dimensions spécifiques aux documents.

3.2 TYPOLOGIE DES MESURES

Dans notre modèle, nous avons associé aux mesures les fonctions d'agrégation compatibles avec l'additivité de celles-ci. Les mesures peuvent être additives, semi-additives ou non-additives [Kimball & Ross, 2002 ; Horner et al., 2004]. Pour répondre aux spécificités des collections de documents, nous définissons une extension du concept classique de mesure. Nous distinguons ainsi deux types de mesures : les mesures numériques et les mesures textuelles. Cette typologie des mesures induit une association de fonctions d'agrégation compatibles.

Une **mesure numérique** est exclusivement composée de données numériques et elle est soit additive, soit semi-additive. Avec une mesure additive, toutes les fonctions d'agrégation traditionnelles peuvent être employées. Avec les mesures semi-additives, seules certaines fonctions d'agrégation peuvent être employées. Par exemple, dans le cadre de l'analyse multidimensionnelle de données documentaires, le nombre de mots clés peut constituer une mesure numérique. Vous trouverez tous les détails de cette modélisation multidimensionnelle de données documentaires dans [Ravat et al. , 2007h].

En plus des mesures numériques classiquement utilisées dans les magasins de données, nous proposons d'intégrer des **mesures textuelles**. Ces mesures étant composées de texte, elles sont systématiquement non-additive. Ce texte peut représenter un mot, un paragraphe ou encore un document complet. Avec une mesure non-additive, seules les fonctions d'agrégation génériques peuvent être employées (telles que COUNT et LIST). Pour contourner cette limite, nous proposons l'utilisation de fonctions d'agrégation spécifiques. Les auteurs de [Park et al., 2005] suggèrent l'emploi de fonctions d'agrégation inspirées de la fouille de texte, telles que TOP_KEYWORDS qui retourne les N principaux mots-clefs d'un document et SUMMARY qui génère le résumé d'un document. De notre côté, nous avons proposé la fonction AVG_KW qui combine plusieurs mots-clefs en un mot-clef plus général [Ravat et al., 2007b].

3.3 DIMENSIONS REPRESENTANT UN DOCUMENT

Du fait des spécificités des données analysées, plusieurs types de dimensions peuvent être identifiés. Afin de préciser le type de données servant de support à l'élaboration des dimensions, nous proposons deux types de dimensions.

1. **Les dimensions de méta-données** représentent les informations relatives au contexte des documents. Par exemple, les auteurs, l'éditeur, la date de publication d'un article voire les méta-données du Dublin Core[11] peuvent être représentés avec ce type de dimension.

2. **Les dimensions documentaires** sont construites à partir du contenu et de la structure logique des documents. La hiérarchisation de ce type de dimension représente la structure logique générique d'une collection de documents. Une constellation textuelle peut contenir un ensemble de dimensions de structures, mais un fait ne peut être relié qu'à une unique dimension structure.

Exemple. Afin d'analyser les données d'un laboratoire de recherche, un utilisateur étudie le contenu d'une collection de documents composée d'articles scientifiques. Ces articles ont un même en-tête et une structure logique plus ou moins complexe. De plus, ces articles contiennent un certain nombre d'informations et de méta-données : noms des auteurs, leur affiliation (institut, pays), la date de publication, une liste de mots-clefs… Pour répondre à ce besoin, nous proposons de construire un schéma multidimensionnel contenant une dimension documentaire (STRCTURE) et trois dimensions de méta-données (MOTS-CLEFS, TEMPS et AUTEUR)

[11] Dublin Core Metadata initiative (DCMI) de http://dublincore.org/

Le schéma multidimensionnel que nous obtenons est le suivant :

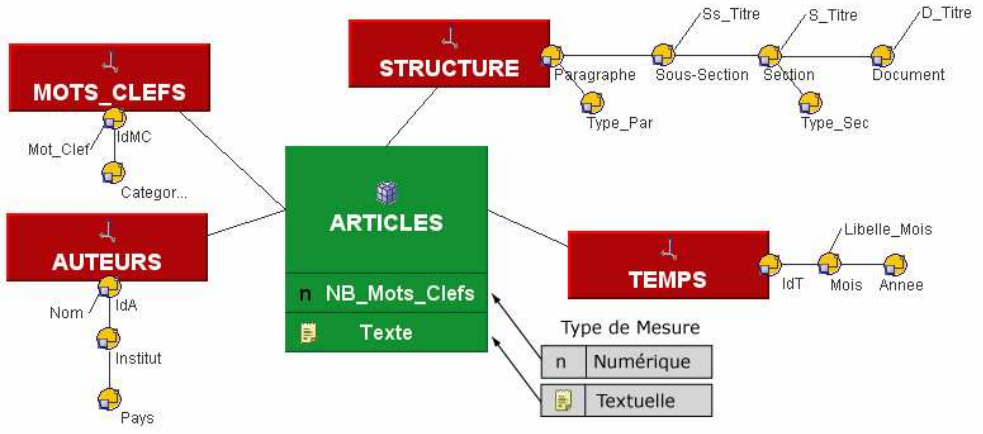


Figure 20 : Schéma multidimensionnel conceptuel de données documentaires

La dimension STRUCTURE est composée de paramètres représentant la structure logique générique des documents. La structure générique définie par la dimension STRUCTURE peut ne pas représenter complètement la structure des documents de la collection, la hiérarchie de cette dimension est une hiérarchie dite « non recouvrante » (« non-covering » ou « ragged » en anglais) [Malinowski & Zimányi, 2006], c'est-à-dire une hiérarchie où tous les paramètres ne sont pas nécessairement instanciés. Par exemple, dans la collection, dans certains articles, il se peut que certains paragraphes soient regroupés en sections mais sans sous-sections.

Au niveau logique, l'intégration de données documentaires dans un schéma multidimensionnel induit une extension des technologies R-OLAP (OLAP relationnel) [Kimball & Ross, 2002]. Cette modélisation logique repose sur deux espaces de stockage : une base de données multidimensionnelles et un espace de stockage de fichiers XML. Dans la base de données multidimensionnelles, chaque dimension est modélisée par une table relationnelle et chaque table de fait contient les mesures et les clés étrangères vers les tables dimensions. Les documents sont stockés à part, dans des fichiers XML. Ils sont reliés aux données factuelles via une expression XPath.

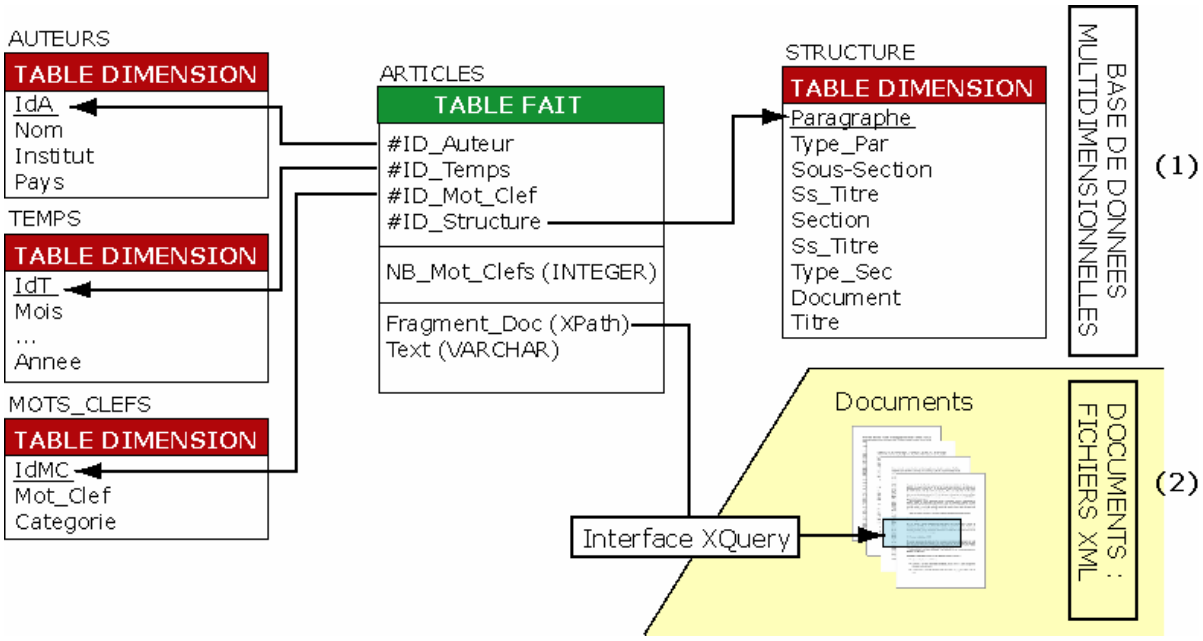


Figure 21 : Schéma logique multidimensionnel de données documentaires

4 GESTION DE LA COHERENCE SEMANTIQUE

Le modèle présenté dans la section 2 propose une représentation conceptuelle des données décisionnelles (constellation de faits reliés à des dimensions multi-hiérarchisées et multi-instanciées) et une visualisation adaptée de ces données (table multidimensionnelle). Or, ce modèle ne tient pas compte de certaines incohérences pouvant survenir lors de l'élaboration, de l'alimentation ou de l'utilisation d'un magasin de données multidimensionnelles. Ne pas prendre en compte ces incohérences peut être préjudiciable pour les prises de décision et donc pour l'avenir de l'organisation considérée.

Dans le cadre de cette section, nous souhaitons proposer une solution pour assurer la cohérence sémantique des données décisionnelles. Cette cohérence vise à offrir un cadre d'analyses fiables interdisant des corrélations incohérentes de données. L'objet de cette section est de vous présenter les résultats que nous avons obtenus dans ce domaine.

4.1 PROBLEMATIQUE

Notre objectif est d'éviter l'utilisation de données incohérentes préjudiciables pour les prises de décisions. Par exemple sur une même dimension géographique, la définition d'une hiérarchie organisant les villes en départements et en régions est cohérente pour des villes françaises tandis que cette hiérarchie propre à la France ne peut être appliquée pour des villes situées dans des pays étrangers. D'autre part, certaines analyses peuvent s'avérer inconsistantes voire inexploitables : par exemple, il est impossible d'analyser les ventes de produits classés selon une taxonomie française dans des magasins classés suivant la géographie américaine (états, villes).

Nos travaux visent à étendre le modèle de base afin qu'il puisse supporter des informations cohérentes et fiables nécessaires à des prises de décisions judicieuses ayant un impact sur l'avenir de l'organisation ou de l'entreprise.

Pour la conception d'un modèle de données décisionnelles fiables et cohérentes, nous proposons d'intégrer des **contraintes** dans le modèle multidimensionnel de base [Samtani et al., 1998 ; Hurtado & Mendelzon, 2002]. La plupart des modèles multidimensionnels s'intéressent aux contraintes structurelles. Ces modèles traitent le problème d'additivité des mesures le long des hiérarchies [Lehner, 1998 ; Hurtado & Mendelzon, 2002 ; Lechtenbörger & Vossen, 2003]. Seuls les modèles de [Carpani & Ruggia, 2001] et de [Hurtado & Mendelzon, 2002] intègrent les contraintes sémantiques. Or, ces travaux se limitent à l'intégration de contraintes pour gérer des incohérences entre les hiérarchies d'une même dimension. Cependant, le langage proposé se focalise sur l'expression de contraintes, et ne permet pas de formaliser les différents types de contraintes. En outre, d'autres conflits peuvent être mis en évidence, notamment entre les hiérarchies de dimensions distinctes.

4.2 TYPOLOGIE DES CONTRAINTES

Le modèle que nous proposons doit assurer la cohérence et la fiabilité des données multidimensionnelles en intégrant un ensemble de contraintes permettant de représenter les règles de gestion de l'organisation étudiée. L'originalité de notre approche réside dans la proposition d'une typologie de contraintes permettant d'identifier clairement les différentes catégories d'incohérences pouvant survenir. Nos travaux présentent l'avantage de non seulement exprimer des contraintes sur les hiérarchies d'une dimension (comme [Carpani et al, 2001] et [Hurtado et al, 2002]), mais également de définir des contraintes inter-dimensions. Nous avons défini deux types de contraintes : structurelles et sémantiques.

Les contraintes structurelles permettent de s'assurer que la structure d'un schéma multidimensionnel est valide. Elles permettent de préciser la bonne utilisation des différents concepts présentés dans la section 2. Ces contraintes portent sur les concepts d'un schéma multidimensionnel. Ces contraintes permettent de vérifier :

- l'unicité des noms de fait, de dimension, de hiérarchie et d'attribut,
- le contenu non vide d'une constellation (elle comporte au moins un fait et une dimension), d'un fait (il comporte au moins une mesure), d'une dimension (elle comporte au moins un paramètre) ou d'une hiérarchie,
- le non-isolement des faits et des dimensions (tout fait est associé à au moins une dimension et vice versa),
- l'acyclicité des paramètres d'une même hiérarchie.

Les contraintes sémantiques sont liées au contexte d'analyse et permettent de préciser certaines règles de gestion. Ces règles sont définies lors de la conception d'un schéma multidimensionnel et permettent d'assurer une alimentation fiable et une manipulation cohérente des données décisionnelles. Pour répondre à ce besoin et en fonction de la portée des contraintes, nous avons défini deux familles de contraintes sémantiques : les contraintes intra-dimension et les contraintes inter-dimensions

Les ***contraintes intra-dimension*** s'appliquent aux hiérarchies de la même dimension et agissent sur les instances de celle-ci. Elles permettent de préciser les relations sémantiques entre les instances d'une même dimension appartenant à deux de ses hiérarchies. Par exemple, une agence de location toulousaine appartient à la hiérarchie décrivant la géographie française et n'appartient pas à la hiérarchie représentant la géographie américaine (cf. dimension Agences en section 2.3). Ces contraintes permettent d'exprimer des règles d'exclusion, d'inclusion, de simultanéité, de totalité et de partition entre les hiérarchies d'une même dimension.

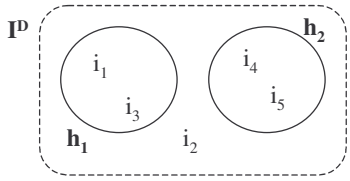
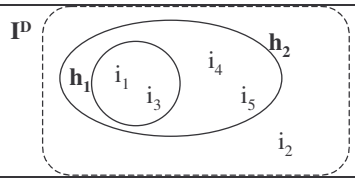
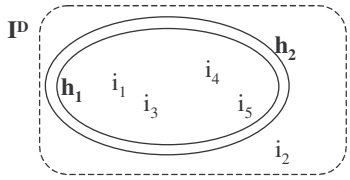
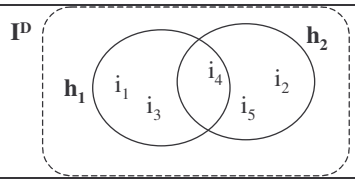
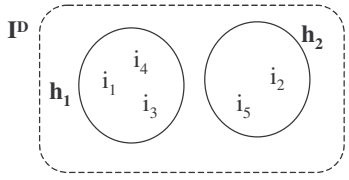
Les ***contraintes inter-dimensions*** s'appliquent aux hiérarchies de différentes dimensions et agissent sur les instances du fait associées à ces dimensions. Elles permettent de spécifier selon quels axes (dimensions) et/ou quelles perspectives (hiérarchies) peuvent être associées les indicateurs d'activité d'un fait (mesures). Ces spécifications caractérisent les relations d'exclusion, d'inclusion, de simultanéité, de totalité et de partition entre les hiérarchies de deux dimensions.

Les sections suivantes proposent une définition précise de ces contraintes sémantiques.

4.3 CONTRAINTES SEMANTIQUES INTRA-DIMENSIONS

A partir des relations pouvant exister entre les instances d'une dimension, nous avons défini 5 contraintes sémantiques intra-dimension. En préambule à leurs définitions, nous posons :

- D une dimension,
- $h1 \in H^D$, $h_2 \in H^D$ deux hiérarchies de D,
- $i_1 \in I^D$ et $i_2 \in I^D$ deux instances de D.

Contraintes	Définitions	Illustrations
Exclusion $h_1 \otimes h_2$	Toute instance de D appartenant à une hiérarchie n'appartient pas à la seconde hiérarchie et réciproquement. $\forall i_1 \in h_1 \wedge \forall i_2 \in h_2 \Rightarrow i_1 \neq i_2$ Notons que $h_1 \otimes h_2 = h_2 \otimes h_1$	
Inclusion $h_1 \circledcirc h_2$	Toutes les instances de D appartenant à h_1 appartiennent à h_2: $\forall i_1 \in h_1 \Rightarrow i_1 \in h_2$ Notons que $h_1 \circledcirc h_2 \neq h_2 \circledcirc h_1$	
Simultanéité $h_1 \ominus h_2$	Toutes les instances de D appartenant à h_1 appartiennent à h_2 et vice versa : $\forall i_1 \in h_1 \Leftrightarrow i_1 \in h_2$ Notons que - $h_1 \ominus h_2 = h_2 \ominus h_1$ - $h_1 \ominus h_2 \Leftrightarrow h_1 \circledcirc h_2 \wedge h_2 \circledcirc h_1$	
Totalité $h_1 \oplus h_2$	Toutes les instances de D appartiennent à h_1 et/ou à h_2: $\forall i_1 \in I^D, i_1 \in h_1 \vee i_1 \in h_2$ Notons que $h_1 \oplus h_2 = h_2 \oplus h_1$	
Partition $h_1 \oslash h_2$	Toute instance de D appartient h_1 ou (exclusif) à h_2: $\forall i_1 \in I^D, (i_1 \in h_1 \wedge i_1 \notin h_2) \vee (i_1 \notin h_1 \wedge i_1 \in h_2)$ Notons que - $h_1 \oslash h_2 = h_2 \oslash h_1$ - $h_1 \oslash h_2 \Leftrightarrow h_1 \otimes h_2 \wedge h_1 \ominus h_2$	

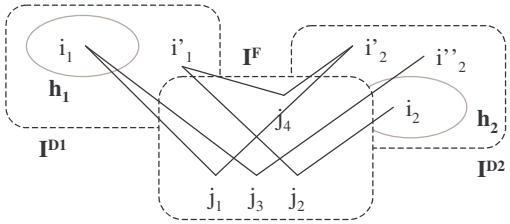
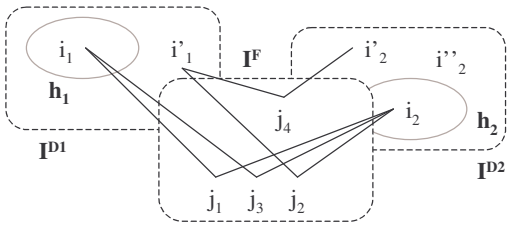
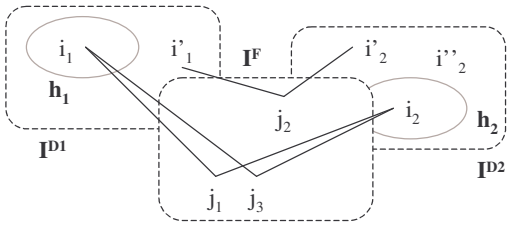
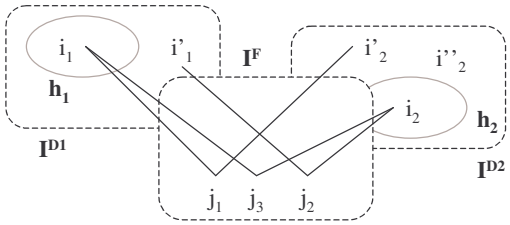
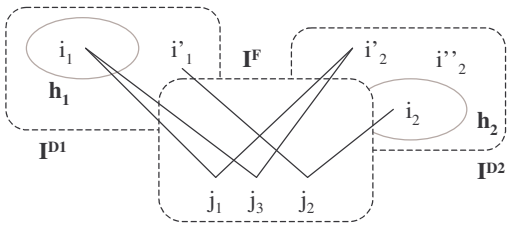
Exemple. Dans la Figure 17, les agences de locations sont situées soit aux Etats-Unis, soit en France. Toute agence située aux Etats-Unis ne peut être analysée selon la hiérarchie de la géographie française, et réciproquement. De plus, toute agence possède une zone géographique dans son pays. Pour compléter la sémantique de la constellation, nous modélisons (sur la dimension *AGENCES*) les contraintes de partition et d'inclusion intra-dimension suivantes :

- $h^{AGENCES}_1 \oslash h^{AGENCES}_2$ (partition),
- $h^{AGENCES}_1 \circledcirc h^{AGENCES}_3$, $h^{AGENCES}_2 \circledcirc h^{AGENCES}_3$ (inclusions).

4.4 CONTRAINTES SEMANTIQUES INTER-DIMENSIONS

Les contraintes inter-dimensions sont exprimées entre les hiérarchies de dimensions distinctes. Ces contraintes décrivent les relations entre les données d'un fait en considérant les perspectives d'analyse appliquées. Il s'agit de contraintes portant sur les instances du fait associées aux sous-ensembles d'instances des hiérarchies des dimensions.

On pose D_1 et D_2 deux dimensions associées à un fait F ($D_1 \in Star^C(F) \wedge D_2 \in Star^C(F)$) et $h_1 \in H^{D1}$, $h_2 \in H^{D2}$ deux hiérarchies des dimensions D_1 et D_2.

Contrainte	Définitions	Illustrations
Exclusion $h_1 \otimes h_2$	Toute instance de F liée à une instance de D_1 appartenant à h_1 ne peut être liée à une instance de D_2 appartenant à h_2 et réciproquement. $\forall j \in I^F, \exists i_1 \in h_1 \mid i_1 \in IStar^F(j) \Rightarrow \neg(\exists i_2 \in h_2 \mid i_2 \in IStar^F(j))$ Notons que $h_1 \otimes h_2 = h_2 \otimes h_1$	
Inclusion $h_1 \circledcirc h_2$	Toutes les instances de F liées aux instances de la dimension appartenant à h_1 sont également liées aux instances appartenant à h_2. $\forall j \in I^F, \exists i_1 \in h_1 \mid i_1 \in IStar^F(j) \Rightarrow \exists i_2 \in h_2 \mid i_2 \in IStar^F(j)$ Notons que $h_1 \circledcirc h_2 \neq h_2 \circledcirc h_1$	
Simultanéité $h_1 \ominus h_2$	Toutes les instances de F liées aux instances de la dimension appartenant à h_1 sont également liées aux instances de h_2 et réciproquement $\forall j \in I^F, \exists i_1 \in h_1 \mid i_1 \in IStar^F(j) \Leftrightarrow \exists i_2 \in h_2 \mid i_2 \in IStar^F(j)$ Notons que $h_1 \ominus h_2 = h_2 \ominus h_1$ et $h_1 \ominus h_2 \Leftrightarrow h_1 \circledcirc h_2 \wedge h_2 \circledcirc h_1$	
Totalité $h_1 \oslash h_2$	Toute instance de F est liée à une instance appartenant à l'une des deux hiérarchies et éventuellement aux deux hiérarchies $\forall j \in I^F, (\exists i_1 \in h_1 \mid i_1 \in IStar^F(j)) \vee (\exists i_2 \in h_2 \mid i_2 \in IStar^F(j))$ Notons que $h_1 \oslash h_2 = h_2 \oslash h_1$	
Partition $h_1 \oslash h_2$	Chaque instance de F est associée soit aux instances de h_1 soit à celles de h_2 (ou exclusif). $\forall j \in I^F, (\exists i_1 \in h_1 \mid i_1 \in IStar^F(j) \wedge \neg\exists i_2 \in h_2 \mid i_2 \in IStar^F(j)) \vee (\neg\exists i_1 \in h_1 \mid i_1 \in IStar^F(j) \wedge \exists i_2 \in h_2 \mid i_2 \in IStar^F(j))$ Notons que $h_1 \oslash h_2 = h_2 \oslash h_1$ et $h_1 \oslash h_2 \Leftrightarrow h_1 \otimes h_2 \wedge h_1 \ominus h_2$	

Exemple. Comme représenté en Figure 17 de ce chapitre, la dimension VEHICULES possède trois hiérarchies. La hiérarchie "$h^{VEHICULES}_2$" intitulée "class-us", spécifique aux Etats-Unis, décrit les véhicules suivant une classification en type de moteur, puis en classe de véhicule (confort, luxe, …), tandis que la hiérarchie "$h^{VEHICULES}_1$" (intitulée "class-fr")décrit une nomenclature inhérente à la France. La hiérarchie "$h^{VEHICULES}_3$" organise les véhicules à louer selon leur genre (sport, citadine, …) et en nombre de vitesse. Pour exprimer le fait qu'une agence française (respectivement américaine) ne propose pas la location de véhicule décrit selon la classification des Etats-Unis (respectivement française), il faut définir les deux contraintes de partition suivantes $h^{AGENCES}_1 \oslash h^{VEHICULES}_2$ et $h^{AGENCES}_2 \oslash h^{VEHICULES}_1$.

Dans la figure suivante, nous représentons la constellation définie en Figure 17 complétée par les deux contraintes inter-dimensions définies au-dessus ainsi que les trois contraintes intra-dimensions définies précédemment.

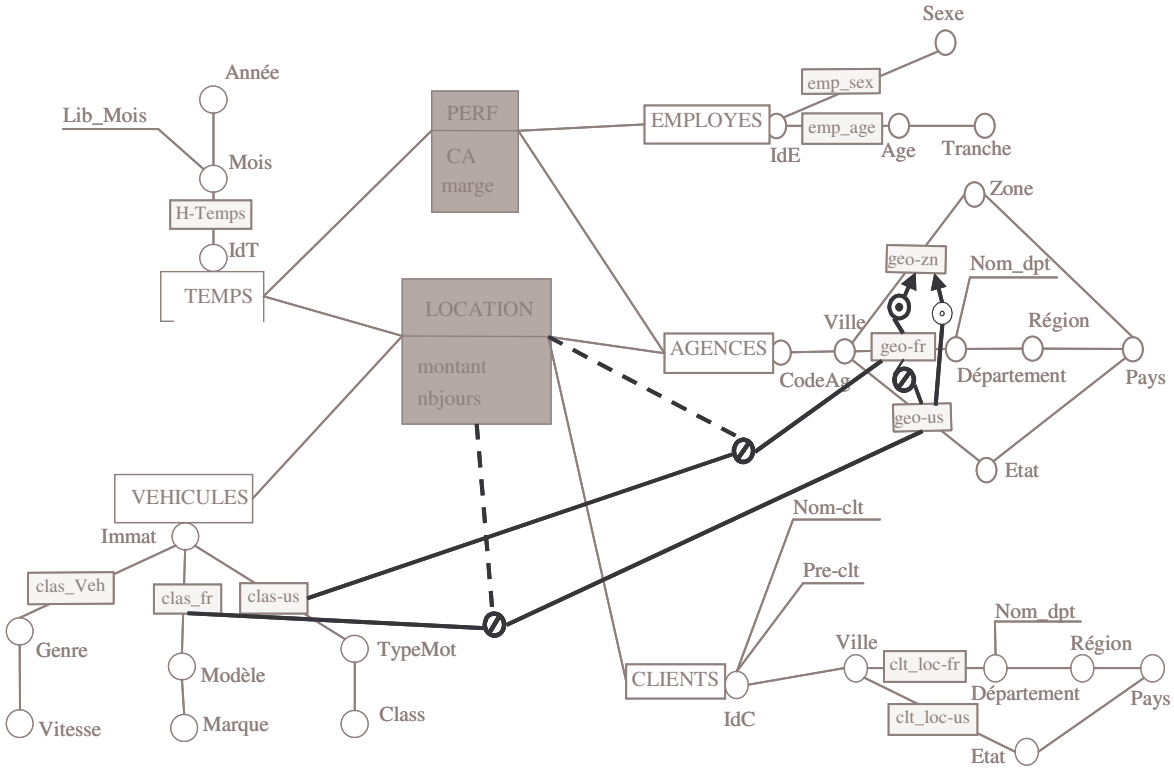


Figure 22 : Exemple complet d'une constellation

L'ensemble des contraintes présentées en section 4 offre une représentation plus précise de la réalité et visent surtout à interdire les corrélations incohérentes lors des analyses. Les contraintes sémantiques expriment l'inclusion, l'exclusion, la simultanéité, la totalité et la partition entre les hiérarchies [Ghozzi et al, 2003b]. Nous distinguons les contraintes intra-dimension qui permettent de caractériser les relations entre les instances des hiérarchies d'une même dimension, des contraintes inter-dimensions qui caractérisent les interactions entre les instances des faits reliées aux instances des hiérarchies de dimensions distinctes.

5 GESTION DE LA COHERENCE TEMPORELLE

Certains auteurs [Vaisman & Mendelzon, 2001] considèrent qu'un fait reflète la partie dynamique d'un schéma multidimensionnel tandis que les dimensions représentent la partie statique. Cependant, les valeurs et les schémas des données sources et par voie de conséquence des entrepôts de données construits à l'aide de ces données peuvent évoluer au cours du temps. De la même manière, les besoins des décideurs peuvent évoluer (suppression de composants inutiles, ajout de nouveaux paramètres, de nouvelles dimensions, de nouveaux faits etc.). Il serait donc pertinent d'intégrer ces changements dans des schémas multidimensionnels. Nous pouvons remarquer que ces changements peuvent intervenir aussi bien au niveau des faits qu'au niveau des dimensions ; autrement dit, ils peuvent porter sur tous les composants d'un schéma multidimensionnel. La gestion de la cohérence temporelle vise à fournir des schémas multidimensionnels pouvant évoluer dans le temps.

5.1 PROBLEMATIQUE

Les travaux relatifs à la prise en compte de l'évolution temporelle des données multidimensionnelles sont à la croisée des travaux sur les bases de données temporelles ou à

versions et les bases de données multidimensionnelles. Durant ces dernières années, de nombreuses propositions ont été faites pour des modèles de données relationnelles ou objet. Nous pouvons notamment citer les travaux de notre équipe pour des bases de données objet à versions. Peu d'équipes de recherche ont apporté une solution pour la gestion de l'évolution de Bases de Données Multidimensionnelles (BDM). Or, de par ses composants spécifiques, la gestion de l'évolution temporelle dans un modèle multidimensionnel constitue un axe de recherches à part entière.

Comme indiqué dans [Body et al., 2002 ; Wrembel & Morzy, 2005], la gestion des changements d'un schéma multidimensionnel peut être classée en deux catégories. La première catégorie permet de gérer les évolutions de données et de valeurs. Ces travaux [Blaschka et al., 1999 ; Hurtado et al, 1999 ; Vaisman et al., 2002] permettent à un décideur de manipuler la dernière version du schéma et de ses instances. Dès qu'un changement survient, le schéma est converti et les états passés sont perdus. La seconde catégorie permet de dater (avec des intervalles de validité) les modifications d'un schéma et de les sauvegarder au travers de versions. La majorité des travaux se sont centrés sur un seul sujet d'analyse - modèle en étoile - (Body et al., 2002 ; Wrembel & Morzy, 2005] et proposent uniquement le changement des instances et des structures des dimensions. Certaines propositions s'avèrent complexes et difficiles à utiliser. Par exemple, dans [Body et al., 2002], une dimension temporelle regroupe l'ensemble des versions de membres estampillées et l'ensemble des relations inter-versions de membres également estampillées.

Notre objectif est de pouvoir gérer de manière aisée mais complète les changements pouvant intervenir dans une constellation telle que nous l'avons définie en section 2.5. Afin de permettre à un décideur de pouvoir manipuler les différentes évolutions d'une BDM, nous proposons d'utiliser un modèle à base de **versions**. Nous souhaitons gérer aussi bien les évolutions de données que les évolutions de structures. Pour les **évolutions de données**, notre modèle doit supporter :

- l'insertion et la suppression d'instances de dimension (afin de modéliser, par exemple, l'ajout d'un nouveau produit ou la suppression d'un produit sorti de la vente),
- la mise à jour d'instances d'attributs d'une dimension (pour modéliser par exemple, le changement de libellé d'un produit).

Pour les **évolutions de structure**, notre modèle doit supporter

- l'ajout ou la suppression d'un fait,
- la modification d'un fait (ajout ou suppression d'une mesure voire d'une relation fait-dimension),
- l'ajout ou la suppression d'une dimension,
- la modification d'une dimension (ajout ou suppression d'une hiérarchie, ajout ou suppression d'un attribut – paramètre ou attribut faible – voire réorganisation des niveaux d'agrégation d'une hiérarchie).

5.2 PRINCIPES

Pour répondre à ce besoin, nous devons étendre le principe de la constellation tel qu'il a été énoncé en section 2.5. Pour gérer l'évolution des données dans un schéma conceptuel multidimensionnel, la constellation devient un ensemble de contextes d'analyse fiable pour une période de temps donnée. Un contexte d'analyse correspond à un sujet d'analyse et ses axes pour une période de temps donnée.

D'un point de vue conceptuel, une constellation devient une collection de versions d'étoiles. Une version d'étoile correspond à un contexte d'analyse fiable à une période de temps donné. Cette version d'étoile contient une version du fait, ses versions de dimension reliées et son intervalle de validité. Lors de l'alimentation du magasin multidimensionnel, si seule une évolution de valeurs intervient, il n'y a pas de changement de version. Aussi, au sein d'une version de fait ou de dimension, les instances sont estampillées. Par contre, si lors de l'alimentation du magasin, un changement de structure est détecté, une nouvelle version de fait ou de dimension est crée. Dès qu'une nouvelle version de fait ou de dimension est créée, une nouvelle version d'étoile est également créée car nous changeons de contexte d'analyse.

La gestion de l'évolution est définie au travers d'un modèle temporel linéaire et discret. Un instant est un point sur la ligne de temps tandis qu'un intervalle représente le temps entre deux instants. Nous considérons les temps de validité et les temps de transactions [Bertino et al., 1996]. Le temps de validité correspond au temps quand l'information est valide dans le monde réel tandis que le temps de transaction correspond au moment où l'information est stockée dans le magasin multidimensionnel. Notons, qu'il existe différents temps de transaction : au niveau des sources, de l'entrepôt puis du magasin. Au niveau du magasin de données multidimensionnelles, chaque extraction fournit un instant dans le temps de transaction.

Les sections suivantes proposent une définition plus explicite des concepts de versions d'étoile, de fait et de dimension.

5.3 CONSTELLATION ET VERSIONS D'ETOILE

Dans cette extension, une constellation est modélisée par un ensemble de versions d'étoile.

Définition. Une constellation C est définie par une collection de versions d'étoile $\{VS_1,\ldots, VS_U\}$ où $\forall i \in [1..u]$, VS_i est définie par $(VF, \{VD_1,\ldots, VD_V\}, T)$

- VF une version de fait,
- $\forall k \in [1..v]$, VD_k est une version de dimension associée à une version de fait,
- $T = [T_{Start}, T_{End}]$ est l'intervalle temporel durant lequel la version d'étoile est valide.

Exemple : le schéma suivant présente les évolutions d'une constellation composée de deux faits et de trois dimensions.

- L'instant T_1 correspond au premier point d'extraction et a permis de construire une constellation composée de deux versions d'étoile ($VS_{1.1}$ et $VS_{2.1}$) correspondant à deux contextes d'analyse. La version d'étoile $VS_{1.1}$ contient une version de fait $VF_{1.1}$ et les premières versions des dimensions D_1 et D_3 (respectivement intitulées $VD_{1.1}$et $VD_{3.1}$). La version d'étoile $VS_{2.1}$ permet d'analyser la première version du fait F2 ($VF_{2.1}$) en fonction des premières versions des dimensions D_3 et D_4 (respectivement libellées $VD_{3.1}$ et $VD_{4.1}$).
- L'instant T_2 correspond à un point d'extraction n'engendrant pas de changement de structure pour les versions de fait ou de dimension définie en T_1.
- L'instant T_3 correspond au point d'extraction engendrant des changements structurels et la création de deux nouvelles versions d'étoile. La première intègre une nouvelle version du fait F_1 car F_1 possède une nouvelle mesure. Cette nouvelle version du fait ($VF_{1.2}$) est maintenant reliée à la dimension D_2 ($VD_{2.1}$) et à une nouvelle version de la dimension D_3. La seconde version d'étoile permet d'intégrer également la seconde version de la dimension D_3.
- Comme l'instant T_2, l'instant T_4 n'engendre pas de changement structurel.

La définition de ces différentes versions d'étoile est effectuée comme suit :

- $VS_{1.1} = (VF_{1.1}, \{VD_{1.1}, VD_{3.1}\}, [T_1,T_3])$
- $VS_{2.1} = (VF_{2.1}, \{VD_{3.1}, VD_{4.1}\}, [T_1,T_3])$
- $VS_{1.2} = (VF_{1.2}, \{VD_{1.1}, VD_{2.1}, VD_{3.2}\}, [T_3,T_{Now}])$
- $VS_{2.2} = (VF_{2.1}, \{VD_{3.2}, VD_{4.1}\}, [T_3,T_{Now}])$

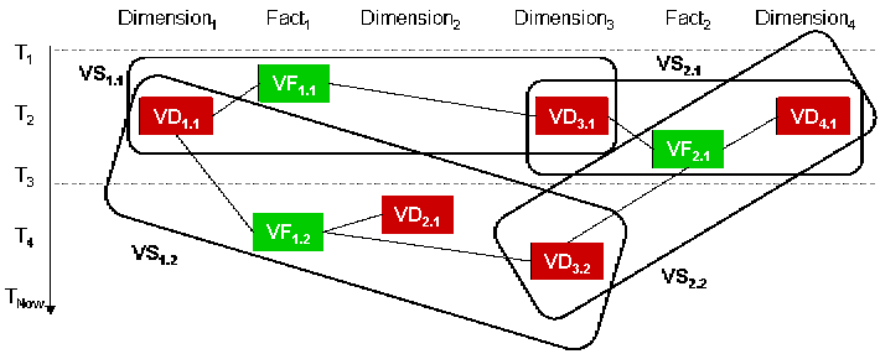


Figure 23 : Exemple de versions d'étoile*Erreur !*

Notre modèle est qualifié de multi-versions car pour un même instant, plusieurs versions d'étoile peuvent être définies. Si le changement des données sources ne nécessite pas de changement structurel au niveau de l'étoile, la version d'étoile courante est rafraîchie : une nouvelle instance de dimension et ou une nouvelle instance de fait est ajoutée. Si l'intégration nécessite un changement de la structure de la version d'étoile (par exemple, le changement d'une hiérarchie dans une dimension), une nouvelle version d'étoile est créée.

5.4 COMPOSANTS D'UNE VERSION D'ETOILE

Chaque version d'étoile est composée d'une version de fait et de plusieurs versions de dimensions (une seule par dimension). Chaque version de dimension est composée de propriétés qui sont organisées selon différentes hiérarchies.

Définition. Une version de fait VF est définie par $(N^{VF}, Int^{VF}, Ext^{VF}, Map^{VF})$

- N^{VF} est le nom du fait
- $Int^{VF} = \{m_1,\ldots, m_p\}$ est l'intention du fait qui correspond à un ensemble de mesures,
- $Ext^{VF} = \{i^{VF}_1,\ldots, i^{VF}_x\}$ est l'extension du fait composée d'instances définies comme suit : $\forall k \in [1..x]$, $i^{VF}_k = [m_1:v_1,\ldots, m_p:v_p, id^{VD_1}:id_1,\ldots, id^{VD_v}:id_v, T_{Start}:vt, T_{End}:vt']$ où $m_1:v_1,\ldots, m_p:v_p$ sont des valeurs de mesures, $id^{VD_1}:id_1,\ldots, id^{VD_v}:id_v$ les identifiants des dimensions liées et $T_{Start}:vt$, $T_{End}:vt'$ sont les valeurs des temps de transaction,
- Map^{VF} est la fonction de mapping permettant d'alimenter la version de fait.

Remarques :

- Toutes les versions de fait possédant le même nom (N^{VF}) correspondent à un fait. Autrement dit, chaque version de fait représente un état durant son cycle de vie.
- Le temps de transaction d'une version d'étoile peut être déduit à partir des intervalles de temps associés aux instances de dimension et de fait qu'elle contient. $\forall i \in [1..u]$, $VS_i,=(VF^i, \{VD^i_1,\ldots, VD^i_V\}, [T_{Start}{}^i, T_{End}{}^i])$, $\forall k \in [1..x]$, $i^{VF}_k \in Ext^{VFi}$, then $T_{Start}{}^i \leq T_{Start}{}^{VFi}_k \wedge T_{End}{}^{VFi}_k \leq T_{End}{}^i \wedge T_{Start}{}^{VFi}_k \leq T_{End}{}^{VFi}_k$.

- De la même manière, nous pouvons calculer les temps de transaction de chaque version de fait ou de dimension voire de chaque fait ou dimension.
- Une nouvelle version de fait est créée lorsqu'une nouvelle mesure est créée ou lorsqu'une ancienne mesure est supprimée.

Définition. Une version de dimension VD est définie par (N^{VD}, Int^{VD}, Ext^{VD}, Map^{VD})

- N^{VD} est le nom de la dimension,
- $Int^{VD} = (A^{VD}, H^{VD})$ est l'intention de la dimension composée d'attributs -$A^{VD} = \{a_1,\ldots, a_q\} \cup \{id^{VD}, All\}$- organisés au travers de hiérarchies -$H^{VD} = \{H^{VD}_1,\ldots, H^{VD}_W\}$-,
- $Ext^{VD} = \{i^{VD}_1,\ldots, i^{VD}_Y\}$ est l'extension de la version de dimension qui est composée d'instances. Chaque instance de dimension est définie comme suit : $\forall k \in [1..Y]$, $i^{VD}_k = [a_1:v_1,\ldots, a_q:v_q, T_{Start}:vt, T_{End}:vt']$ où $a_1:v_1,\ldots, a_q:v_q$ sont des valeurs d'attributs de dimension et $T_{Start}:vt$, $T_{End}:vt'$ sont des valeurs de temps de transaction,
- Map^{VD} est la fonction de mapping pour la population de la version de dimension.

Remarques

- La hiérarchie ne nécessite pas de nouvelle définition.
- Une dimension est définie par plusieurs versions de dimensions ayant le même nom. Une nouvelle version de dimension est définie lors de la modification de sa structure (ajout ou suppression d'un attribut de dimension voire ajout, suppression ou modification d'une hiérarchie [Eder et al., 2001]).

Les fonctions de mapping reposent sur une algèbre et sont définies dans (Ravat et al., 2006a). Elles permettent de définir le processus d'alimentation d'une version (de fait ou de dimension) soit directement à partir des données d'un entrepôt sources, soit à partir d'une version antérieure. Ce principe est schématisé dans la figure suivante :

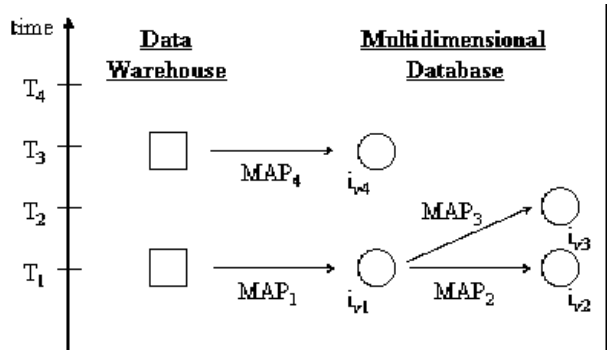


Figure 24 : Principe des fonctions de mapping

6 INTEGRATION ET CAPITALISATION DE L'EXPERTISE DES DECIDEURS

Face à la mondialisation et à la concurrence exacerbée de ces dernières années, les entreprises souhaitent disposer d'outils facilitant les prises de décisions rapides et fiables nécessaires à leur pérennisation et à leur expansion. Dans les sections précédentes, nous avons abordé la gestion de la cohérence temporelle et sémantique afin d'offrir des données fiables et des analyses cohérentes. Or, les données brutes d'une BDM peuvent parfois s'avérer complexes et difficiles à interpréter. En effet, les prises de décisions reposent non seulement sur les données brutes mais également sur les réflexions, les commentaires des analystes voire la confrontation de différentes interprétations. À notre connaissance, il n'existe pas de modèle et de logiciel

permettant aux décideurs d'analyser les données décisionnelles en intégrant les tâches qu'ils conduisent de manière manuelle lors de la prise de décision. En effet, outre l'expression de la requête et la visualisation des données supportées par les logiciels actuels, les analystes ont besoin d'interpréter ces données pour communiquer leurs conclusions et éventuellement en débattre collégialement. Pour répondre à ce manque, nous proposons d'intégrer au sein d'un SAD, ces tâches du processus de prise de décision. Ainsi, mémoriser et réutiliser l'expertise des analystes permettra à l'organisation de préserver le patrimoine immatériel décisionnel tout aussi important que les données elles-mêmes.

6.1 PROBLEMATIQUE

Notre objectif est de proposer une Mémoire d'Expertise Décisionnelle (MED) permettant de modéliser dans un même espace les données décisionnelles à structure multidimensionnelle et l'expertise des analystes. Pour répondre à l'ensemble des besoins, l'expertise décisionnelle doit pouvoir être formulée sur l'ensemble des composants multidimensionnels : schéma ou table multidimensionnels.

L'intégration de l'expertise des décideurs au niveau d'un schéma multidimensionnel permettra d'expliciter la sémantique des composants et des instances d'une BDM. L'intégration de ces informations essentielles pour le processus de prise de décision n'est pas pris en compte dans les travaux existants.

Au niveau des analyses décisionnelles exprimées au travers de TM, nous souhaitons faciliter la tâche des décideurs (usage personnel) tout en permettant le partage d'expertises (usage collectif) :

- **Pour un usage personnel.** La spécification d'analyses décisionnelles au travers d'une TM accompagnée d'une réflexion critique s'apparente au concept de lecture active [Adler & Van Doren, 1972]. Les annotations aident le lecteur à matérialiser sa réflexion. Or, avec les outils actuels, la formulation, la sauvegarde et la restitution informatique d'une TM et de ses annotations n'est pas possible. Par conséquent, les expertises ne peuvent pas être partagées et exploitées, en particulier, la réutilisation est impossible.
- **Pour un usage collectif.** De nos jours, un outil d'analyse mono-utilisateur n'est plus suffisant. En effet, lorsque l'analyse est complexe, l'avis d'un autre expert est souvent sollicité, ce qui peut donner lieu éventuellement à des débats argumentés visant à atteindre un consensus pour une prise de décision collégiale. Malheureusement, le support de cet échange n'est pas assuré par l'application informatique : l'analyste doit utiliser un moyen de communication classique (réunion, appel téléphonique, courrier électronique, etc.) pour établir le contact. Il doit ensuite restituer le contexte complet de l'analyse et formuler sa question à son interlocuteur. Cet échange gagnerait à être informatisé : une discussion en contexte pourrait accélérer la prise de décision (car le contexte n'a plus à être explicitement formulé). De plus, les arguments de la discussion pourraient être sauvegardés et réutilisés lors d'une expertise future.

6.2 PRINCIPES

En réponse à ces besoins, nous proposons d'intégrer des annotations dans une BDM. À notre connaissance, les systèmes d'annotation couplés aux BDM n'ont pas fait l'objet d'étude. Une première solution consiste à transposer le principe des commentaires associés aux schémas de BD transactionnelles aux BD décisionnelles. Par exemple, la commande COMMENT d'Oracle permet d'associer un commentaire à une table, une vue ou une colonne et ce dernier est stocké dans le dictionnaire de données d'Oracle. Cette proposition est insuffisante pour répondre à notre problématique car elle reste difficilement exploitable et très limitée.

Par ailleurs, les annotations ont largement été étudiées dans le contexte de la gestion électronique de documents [Wolfe, 2002]. Elles sont qualifiées "[d'] informations ou marques supplémentaires apportées au document pour l'enrichir avec des explications brèves et utiles afin de permettre au lecteur de conserver une trace de ses réactions et par la suite marquer les passages mis en valeur[12]". Ces annotations sont dites informelles, contrairement aux annotations formelles qui sont employées pour cataloguer et indexer les documents dans des langages formels, avec un vocabulaire strict qui peut être par exemple issu d'une ontologie dans le cadre du Web Sémantique. Dans notre proposition, nous considérons uniquement les annotations informelles car nous ne souhaitons pas contraindre les décideurs à l'utilisation d'un vocabulaire normé et limité.

La majorité des systèmes d'annotation informatisés (SAI) se basent sur des documents électroniques de type textuel. Dans ce cadre, les annotations sont matérialisées sous différentes formes : contenus textuels, marques libres (astérisques, accolades, flèches, soulignement, etc.). Elles permettent principalement de mettre en valeur des passages textuels en y associant éventuellement un commentaire [Marshall, 1998]. À ce jour, on dénombre plus de vingt SAI : ce sont des prototypes de recherche *e.g.* Annotea/Amaya du W3C [Kahan et al., 2002] comme des produits commerciaux tels qu'iMarkup Client ou Microsoft Office Web Discussions [Brush, 2002]. Fondamentalement, un SAI permet de visualiser des documents, d'en consulter toutes les annotations et d'en créer de nouvelles. Nous proposons d'adapter la pratique d'annotations papier sur les éléments des BDM (schéma et TM).

6.3 LES ANNOTATIONS DECISIONNELLES

Comme indiqué précédemment, notre proposition consiste à modéliser le savoir-faire des décideurs au travers **d'annotations**. Les annotations visent à conserver les commentaires formulés lors des analyses et les décisions prises. L'expertise que véhiculent ces annotations est utilisée à des fins personnelles ou collectives et elles contribuent à améliorer les analyses futures. Les annotations décisionnelles contiennent des informations subjectives et objectives.

Définition. Une annotation décisionnelle AD est définie par le couple (IS, DO)

- IS = un ensemble d'informations subjectives regroupant
  - le contenu textuel saisi par le décideur qui annote,
  - le type de l'annotation caractérisant son contenu (commentaire, question, réponse à une annotation, conclusion),
  - la portée (locale à un contexte d'analyse ou globale).
- DO = un ensemble de données objectives comportant
  - son identifiant,
  - sa date de création permettant de caractériser sa position dans le fil de discussion ordonné chronologiquement,
  - son créateur (décideur) avec sa fonction et ses coordonnées,
  - son éventuelle référence à une annotation père (dans un fil de discussion),
  - son point d'ancrage spécifiant la localisation précise de l'annotation.

[12] cf. le site Web du projet Annotea initié par le W3C : http://www.w3.org/2001/Annotea/

6.4 ANCRAGE D'ANNOTATIONS DECISONNELLES

Nous proposons une technique d'ancrage unifiée pour exprimer des annotations aussi bien sur les composants d'un schéma multidimensionnel que sur une TM. Toute annotation globale sera visible lors des différentes manipulations réalisées par les décideurs. Par opposition, une annotation locale est uniquement visible dans son contexte (TM dans laquelle elle a été définie).

La définition EBNF (Extended Backus–Naur Form) du point d'ancrage est la suivante :

Définition. Un point d'ancrage α est défini par le triplet (S, D_1, D_2) où :

- S = {C | TM}[.F[.*f*(m)[=val]?]?]? désigne un ancrage relatif au fait F,
- $D_1 = \lambda$ | D[.H[/p[=pos]?]*]? désigne un ancrage relatif à une dimension,
- $D_2 = \lambda$ | D[.H[/p[=pos]?]*]? désigne un ancrage relatif à une dimension.

Notons que C désigne une constellation, TM désigne une table multidimensionnelle, F est un fait, *f*(m) est une mesure associée à une fonction d'agrégation, val représente une valeur prise par la mesure, D désigne une dimension, H désigne une hiérarchie, p désigne un paramètre, pos représente une valeur prise par le paramètre. De même, [elt]? désigne elt comme optionnel dans la spécification, [elt]* désigne une répétition 0 à n de elt, { elt1 | elt2 } désigne l'alternative entre elt1 et elt2, et λ désigne une expression de chemin vide.

6.4.1 Annotation d'un schéma multidimensionnel

Ces annotations sont définies sur un élément conceptuel d'une constellation (fait, mesure, dimension, hiérarchie, paramètre, attribut faible). Elles sont indépendantes de tout contexte d'analyse, et apparaissent dans toutes les tables multidimensionnelles construites à partir de la constellation. Le point d'ancrage α d'une annotation est défini par le triplet (S, D_1, λ).

Exemple. L'exemple suivant présente la constellation permettant d'analyser les importations d'une société.

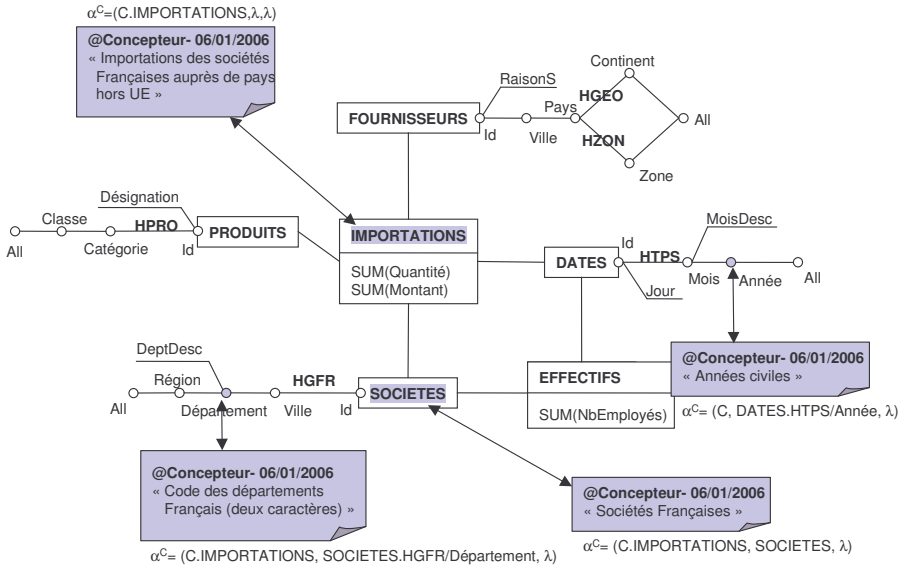


Figure 25 : Exemple d'annotations avec leurs points d'ancrage

Le concepteur du schéma a placé diverses annotations globales,

- au niveau du fait IMPORTATIONS,
- au niveau de la dimension SOCIETES,

- au niveau du paramètre Département (le concepteur a précisé le fait dans le point d'ancrage pour rendre l'annotation disponible uniquement à partir du fait IMPORTATION)
- au niveau du paramètre Année (le fait n'étant pas spécifié dans le point d'ancrage, l'annotation concerne tous les faits liés à la dimension DATES).

6.4.2 Annotation d'une Table Multidimensionnelle (TM)

Ce type d'annotation peut être attaché au fait, aux mesures, à l'une des dimensions ou des hiérarchies courantes (en ligne ou colonne), aux paramètres affichés, aux positions de ces derniers et enfin, aux valeurs observées contenues dans les cellules de la table multidimensionnelle.

Exemple. L'exemple suivant présente la table multidimensionnelle des importations annuelles de produits de la classe 'Electronique' par pays d'origine des fournisseurs. Trois annotations locales à la TM sont disponibles, écrites par le Concepteur, le Directeur des Importations et le Directeur Général.

- L'annotation du concepteur, portant sur le paramètre Année, est une annotation globale spécifiée sur la constellation, et visible sur toutes les tables multidimensionnelles intégrant le paramètre Année.
- L'annotation du Directeur Général porte sur toutes les importations de produits provenant des Etats-Unis (TM.IMPORTATIONS.SUM(Montant), FOURNISSEURS. HGEO/Continent='Amérique'/Pays='Etats-Unis', λ).
- L'annotation du Directeur des Achats porte sur la valeur 230 de la somme des montants des importations (TM.IMPORTATIONS.SUM(Montant)=230) de l'année 2005 (DATES.HTPS/Année=2005) et des fournisseurs chinois (FOURNISSEURS.HGEO/Continent='Asie'/Pays='Chine'). Cette annotation est ancrée dans la TM à partir de la dimension des Fournisseurs et de la dimension des Dates.

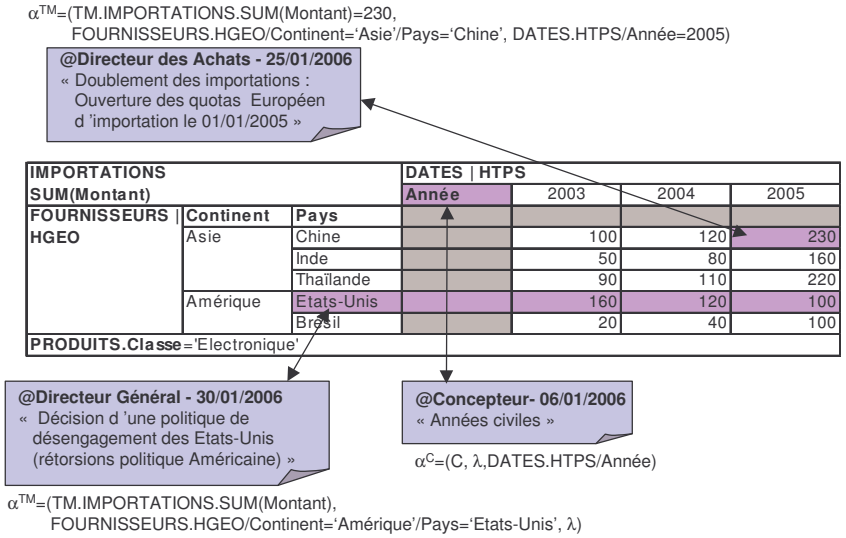


Figure 26 : Exemple d'annotations sur une table multidimensionnelle

7 PERSONNALISATION DE MAGASINS DE DONNEES

A l'heure actuelle, la diversité des besoins d'analyse est satisfaite dans les systèmes décisionnels classiques par la mise en place de divers magasins orientés sujets. Or, la conception d'un MD multidimensionnel est une tâche complexe et le plus souvent assez longue qui à l'heure

actuelle ne repose pas sur des concepts, des formalismes, une démarche et une méthode standards [Rizzi et al., 2006]. De plus, l'implantation d'un système décisionnel composé de nombreux magasins de données multidimensionnelles nécessite la mise en place de processus lourds d'élaboration, d'alimentation et de rafraîchissement des données sans oublier des efforts de maintenance importants. Il est alors difficile d'envisager un magasin de données multidimensionnelles pour chaque décideur. Les systèmes actuels s'avèrent imparfaits, voire inadaptés à ces exigences d'**adaptation**.

7.1 PROBLEMATIQUE

Notre problématique consiste à proposer un système décisionnel personnalisable pour chaque décideur. En fait, nous souhaitons proposer un modèle de MD multidimensionnel configurable pour répondre aux besoins des décideurs.

L'idée de développer des mécanismes permettant de personnaliser un système informatique n'est pas une idée nouvelle. Notamment, nous pouvons citer les travaux du domaine de la Recherche d'Information (RI) [Korfhage, 1997 ; Bouzeghoub & Kostadinov, 2005]. La personnalisation en RI consiste généralement à définir des profils utilisateurs (ensembles plus ou moins structurés de caractéristiques) qui sont exploités aux différentes étapes du processus de RI : indexation, recherche…

La personnalisation de BD décisionnelles a fait l'objet d'une première proposition [Bellatreche et al., 2005] visant à fournir aux décideurs la visualisation la plus adaptée à leurs requêtes. Pour ce faire, les auteurs proposent des contraintes de visualisation précisant la structure du cube de données résultat et des préférences utilisateurs définies à l'aide du concept de pré-ordre total [Koutrika & Ioannidis, 2004] portant sur un attribut d'une dimension. Cette proposition se limite à des cubes de données dont les dimensions contiennent un seul attribut et à l'opération de sélection des données pour calculer le cube résultat.

Notre objectif est de proposer pour chaque décideur un environnement personnalisé reposant sur un modèle de données multidimensionnelles. Cette personnalisation permet de mettre en valeur les données essentielles tout en facilitant leur manipulation lors des analyses décisionnelles. Pour ce faire, nous souhaitons proposer une solution plus complète que celle de [Bellatreche et al., 2005]. Cette personnalisation va permettre d'associer un poids ou une préférence à tout composant d'une BDM voire aux valeurs de ces composants. Nous souhaitons également proposer un langage de personnalisation reposant sur le concept des règles actives. Cette personnalisation doit impacter aussi bien l'affichage final des données décisionnelles que leurs analyses au travers des opérations de forages et de rotation de cubes de données (changement d'axes d'analyse).

7.2 NOS RESULTATS

Pour répondre à notre objectif, nous souhaitons offrir les moyens à un utilisateur de paramétrer le schéma de la constellation de sorte à personnaliser son exploitation. Plus précisément, la personnalisation de la constellation s'opère au niveau des attributs de celle-ci, à savoir les paramètres, les attributs faibles et les mesures. En fonction des paramétrages effectués, les différents attributs sont alors utilisés ou non de manière prioritaire par le système lors de la manipulation de la constellation par l'utilisateur.

Nous proposons deux approches pour personnaliser une constellation. L'approche "**naïve**" consiste à associer un poids aux attributs, fixant ainsi un ordre de priorité d'affichage. Le système décide de l'utilisation d'un attribut en fonction de l'importance qui lui a été donnée. L'approche "**avancée**" consiste à définir un ordre de priorité d'un attribut en fonction du contexte de son utilisation, c'est-à-dire en fonction des opérations de manipulation et des données manipulées. Cette approche repose sur un paramétrage ECA [Widom & Ceri, 1996] de la constellation.

7.2.1 Approche "naïve"

Cette approche consiste à associer un poids à chaque attribut a_i de la constellation. Ce poids, noté w_i, modélise l'importance que l'utilisateur souhaite associer à ai. Afin de faciliter son exploitation, chaque poids est normalisé $0 \leq w_i \leq 1$. Lors de l'exploitation des données personnalisées, le système décisionnel affiche uniquement les données relatives aux attributs ayant un poids supérieur à un seuil fixé. Cette solution offre une grande flexibilité. Néanmoins, les autres attributs dont le poids est inférieur au seuil fixé restent accessibles lors des analyses décisionnelles. Les utilisateurs doivent le préciser explicitement. Ainsi, contrairement aux systèmes traditionnels où le paramètre de granularité maximale est initialement affiché et le décideur doit effectuer des opérations de forage pour afficher les autres paramètres, notre personnalisation affiche directement les paramètres possédant les poids les plus élevés.

7.2.2 Approche "avancée"

Cette seconde approche consiste à personnaliser la constellation en intégrant le contexte d'utilisation de ses attributs. Pour cela, nous proposons un langage de configuration qui repose sur un langage de type ECA (Evènement – Condition – Action).

La commande de définition d'une règle de configuration de la constellation repose sur la syntaxe la suivante :

```
CREATE RULE <nom_règle> ON { <ND> | <NhDi> | <NF> }
WHEN <manipulation>
[IF <condition>]
THEN <action>;
```

Les règles peuvent être associées soit à un fait (NF), soit à une dimension (ND), soit à une hiérarchie (N_hD_i).

`<manipulation>` : cette expression détermine le contexte de la manipulation déclenchant la règle. Nous définissons ce contexte par rapport aux opérations de manipulation qui sont appliquées sur les composants de la constellation.

- `DISPLAYED`
- `ROTATED [FROM` ND_{old}`] [TO` ND_{new}`],`
- `DRILLED-DOWN [ON` $ND_{current}$ `[TO` p_{min}`] [ACCORDING TO` $NHD_{current}$`]]`
- `ROLLED-UP [ON` $ND_{current}$ `[TO` p_{max}`] [ACCORDING TO` $NHD_{current}$`]]`

`<condition>` : les règles peuvent être mises en place de manière conditionnelle. Une condition référence l'état courant de la constellation. Nous introduisons une fonction `current(E) : boolean` permettant de déterminer si un élément E de la constellation est en cours de manipulation ; E∈ {ND, ND.N_hD_i, ND[.N_hD_i].p_k, NF, NF.f_i, NF[.f_i].m_k}.

`<action>` : les actions s'appliquent sur les éléments constitutifs de la constellation. Nous proposons une procédure prédéfinie `priority(E,` w_i`)` permettant d'associer contextuellement un poids à chaque attribut (mesure, paramètre, attribut faible). L'élément E∈ {ND, ND.N_hD_i, ND[.N_hD_i].p_k, NF, NF.f_i, NF[.f_i].m_k} ; lorsqu'il s'agit d'une dimension ND, le poids wi est affecté à tous les attributs de la dimension, lorsqu'il s'agit d'une hiérarchie ND.N_hD_i, le poids w_i est affecté à tous les attributs de la hiérarchie… De manière analogue aux conditions, dans notre contexte de mise en œuvre R-OLAP, l'utilisateur peut définir ses propres actions sous forme de procédures stockées.

Exemple. Notre objectif est de personnaliser la dimension "TEMPS" proposé dans la Figure 17, en fonction du contexte de son utilisation. La règle suivante précise que dans le contexte de l'analyse des performances (clause IF) seront utilisés prioritairement les attributs Année et MoisN. Les trimestres ne seront affichés que si l'utilisateur exprime explicitement ce besoin. En outre, remarquons que cette personnalisation est valable dans le cadre d'opérations d'affichage et de rotation.

```
CREATE RULE display_temps_ventes ON Temps
WHEN displayed OR rotated
IF current(PERF)
THEN priority(Temps.H-Temps.Année,1),
     priority(Temps.H-Temps.Trimestre,0),
     priority(Temps.H-Temps.MoisN,1);
```

8 BILAN ET PERSPECTIVES

Les travaux présentés dans ce chapitre se situent dans le cadre de magasins de données multidimensionnelles. Historiquement, la première catégorie, intitulée "Modèle Cube" [Vassiliadis & Sellis, 1999] permettait de représenter les données d'un sujet d'analyse en tant que cellules d'un cube dont les arrêtes regroupent les valeurs des paramètres des dimensions. Ces travaux présentaient l'inconvénient de ne pas séparer les données de la structure et de ne pas représenter la structure arborescente des dimensions. Afin de pallier ces inconvénients, est apparue la seconde catégorie intitulée "**modèle multidimensionnel**". Ce modèle, sémantiquement plus riche, permet de définir précisément les différents composants d'un schéma multidimensionnel tels que les faits et les dimensions. **Nous situons nos travaux dans cette seconde catégorie.** Cependant, les travaux de cette seconde catégorie manquent de formalisation précise, stable et reconnue par l'ensemble de la communauté scientifique [Rizzi et al., 2006 ; Niemi, et al. 2003]. En réponse à ces différentes lacunes, notre souhait est d'apporter des solutions pour une **modélisation multidimensionnelle orientée décideurs.**

8.1 BILAN SUR LA MODELISATION DES MAGASINS

Dans un premier temps, nous avons proposé des concepts et des formalismes graphiques inhérents à la modélisation multidimensionnelle (dualité fait-dimension, **constellation**, **dimension multi-hiérarchisée**…) [Ravat et al., 2007a]. Ce modèle conceptuel générique présente l'avantage d'intégrer l'ensemble des concepts d'un schéma multidimensionnel tout en proposant une structure de visualisation (**table multidimensionnelle**) adaptée aux décideurs. Même si au début de nos recherches, ces concepts n'étaient pas unanimement reconnus, de nos jours, ils tendent à être couramment utilisés par les chercheurs de la communauté du décisionnel.

Dans un second temps, nous avons proposé un ensemble d'extensions à ce modèle de base afin de prendre en compte différentes spécificités des besoins décisionnels. Plus particulièrement, nous avons voulu répondre aux besoins des décideurs en terme :

- d'intégration de données textuelles dans un contexte OLAP,
- de cohérences sémantique et temporelle des données,
- d'intégration d'expertises décisionnelles,
- de personnalisation des analyses.

La première spécificité que nous avons abordée est l'intégration de données documentaires dans l'analyse OLAP [Ravat et al., 2007c]. Pour répondre à ce besoin, en plus des mesures quantitatives, nous avons proposé l'intégration de mesures qualitatives telles que des **mesures**

textuelles. Une telle spécificité introduisait une nouvelle problématique de recherche, à savoir, proposer des fonctions d'agrégation adaptées à ce type de mesures. Même si certaines propositions intégraient le concept de mesure textuelle et de fonction d'agrégation de base (count), aucune proposition n'offrait de fonction d'agrégation compatible avec l'additivité d'une mesure textuelle. Pour répondre à ce besoin, nous avons proposé la **fonction AVG_KW** qui combine plusieurs mots-clés en un mot-clé plus général en se basant sur une ontologie de mots clés. Cette fonction d'agrégation repose sur un algorithme de parcours d'arbres ontologiques [Ravat et al., 2007b]. De plus, l'intégration de données documentaires introduit la définition de dimensions spécifiques telles que des **dimensions de méta-données** ou des **dimensions combinant contenu et structure logique** [Ravat et al., 2007h].

La seconde extension a pour objectif de répondre à un besoin vital en décisionnel, à savoir, **disposer d'informations fiables**. Notre solution repose sur la technique des contraintes. Les contraintes proposées visent à interdire les corrélations incohérentes lors des analyses. Nous avons proposé un ensemble de **contraintes inter et intra-dimensions** [Ravat et al., 2005a] exprimant l'inclusion, l'exclusion, la simultanéité, la totalité et la partition entre les instances [Ghozzi et al., 2004]. Ces définitions conceptuelles ont été implantées dans un contexte R-OLAP et permettent de réduire significativement le nombre de vues matérialisées utilisées lors des analyses OLAP [Ghozzi et al., 2004]. Notamment, nous avons montré comment les contraintes d'exclusion peuvent servir à éliminer des combinaisons incohérentes d'attributs pour réduire la taille du treillis des vues candidates à la matérialisation [Ghozzi, 2004].

La troisième extension a pour objectif de prendre en compte l'évolution dans le temps des schémas multidimensionnels afin de refléter les changements des besoins des décideurs et des sources. La gestion de cette **cohérence temporelle** s'est traduite par l'extension du concept de constellation afin de supporter des versions d'étoile. Une **version d'étoile** correspond à un contexte d'analyse à une période donnée. Par conséquent, elle regroupe une version de fait et ses versions de dimensions associées [Ravat & Teste, 2006]. Chacune de ces versions contient non seulement la structure de données mais également les primitives algébriques permettant d'alimenter les différentes versions [Ravat et al., 2006a]. Ces propositions doivent être complétées par la définition d'un langage de manipulation adaptée.

Dans ce chapitre, nous avons également présenté les principes relatifs à **l'intégration et à la capitalisation de l'expertise des décideurs** dans un modèle multidimensionnel. Afin de faciliter l'interprétation des données OLAP voire de partager des réflexions, nous avons intégré dans le modèle multidimensionnel le concept d'**annotation**. Ces annotations permettent à un décideur de faire une lecture active de ses données ou un usage collectif avec des fils de discussion [Cabanac et al., 2006a]. Afin de simplifier la gestion des annotations, nous avons proposé un mécanisme d'ancrage unifié des annotations sur les différents composants d'un schéma multidimensionnel ou d'une TM [Cabanac et al., 2006b]. Ces propositions conceptuelles ont été implantées dans un environnement R-OLAP. Les annotations sont définies au travers d'une table dans la métabase décrivant ses différentes caractéristiques (identifiant, méta-concept annoté, auteur, date, identifiant de l'annotation parent…) [Cabanac et al., 2007].

La dernière extension permet de faciliter les prises de décisions en proposant un **environnement personnalisé** à chaque décideur. Cette personnalisation permet de mettre en valeur les données essentielles tout en facilitant leur manipulation lors des analyses décisionnelles. En complément des travaux de [Bellatreche et al., 2005], nous associons un poids (préférence) aux composants d'une BDM voire aux valeurs de ces composants [Ravat et al., 2007f]. Pour plus de flexibilité, nous avons proposé un langage ECA pour la définition de la personnalisation. Cette proposition conceptuelle est complétée par une implantation dans un contexte R-OLAP. Dans ce contexte, l'implantation des préférences s'effectue au travers de méta-tables consultées lors de chaque manipulation [Ravat et al., 2007f].

8.2 PRODUCTION SCIENTIFIQUE

Le travail présenté dans ce chapitre a donné lieu à plusieurs encadrements ou co-encadrements. Notamment, les travaux relatifs à la modélisation des données multidimensionnelles sous contraintes sémantiques se sont concrétisés par le co-encadrement (à 80%) de la thèse de Faiza Ghozzi [Ghozzi, 2004]. Les travaux relatifs à la modélisation multidimensionnelle de données documentaires est étudiée en détails dans la thèse de Ronan Tournier [Tournier, 2007] (co-encadrement à 40%). Nos travaux relatifs à la modélisation de magasins de données multidimensionnelles se sont traduits par l'encadrement ou le co-encadrement de 6 étudiants en Master recherche 2IH de l'Université Paul Sabatier (Toulouse III). Mourad Kaddes a permis d'implanter et de valider nos propositions sur les contraintes sémantiques [Kaddes, 2005]. L. Benakezou et E. Nègre ont permis de valider nos propositions sur les modèles à versions [Benakezou, 2006] et la définition d'espaces temporels dans un schéma multidimensionnel [Nègre, 2005]. La gestion de la politique d'accès aux constellations de données a été proposée dans [Sallami, 2004] ; ces travaux ont été les prémisses à nos propositions relatives à la personnalisation. Notre modèle avec annotations a été implanté par H. Jerbi [Jerbi , 2007]. L'intégration de données documentaires dans un schéma multidimensionnel a été traitée par C. Koussa [Koussa, 2007]. De plus, ces travaux sont également le fruit de l'encadrement d'un magistère SIC (Système d'Information et de Communication) de l'INI (Institut National de Formation en Informatique d'Alger) sur l'intégration de documents XML dans un schéma multidimensionnel (A. Nassim).

Ces travaux se sont déroulés dans le cadre de l'action spécifique CNRS-STIC "GafoDonnées" dont l'objectif était de rassembler les communautés des bases de données, de l'apprentissage automatique, de l'analyse de données et des interfaces homme-machine [Laurent et al., 2002]. Certaines propositions sont le fruit de cette collaboration. De même, ces travaux se sont déroulés dans le cadre de l'action incitative EVOLUTION. Cette action visait à proposer des solutions pour le développement de SAD : architecture, modélisation conceptuelle, logique et physique d'entrepôt et de magasins de données multidimensionnelles [Evolution, 2001]. Ces deux collaborations sont explicitées dans le chapitre VI.

D'un point de vue publications, nous pouvons citer les références suivantes[13] :

- 1 article dans la revue internationale IRECOS [Ravat & Teste, 2006],
- 1 article dans l'ouvrage international "Database Modeling for Industrial Data Management " [Ravat et al., 2005a],
- 5 articles dans des conférences internationales : ER'07 [Ravat et al., 2007e], DAWAK'07 [Cabanac et al., 2007], SEKE'07 [Ravat et al., 2007c], DAWAK'06 [Ravat et al., 2006a], ICEIS'03 [Ghozzi et al., 2003a],
- 2 articles dans une revue nationale : RSTI-ISI [Ghozzi et al., 2004], Document Numérique [Ravat et al., 2007h],
- 6 articles dans des conférences nationales : EDA'07 [Ravat et al., 2007g], INFORSID'07 [Ravat et al., 2007f], EDA'06 [Cabanac et al., 2006a], EGC'06 [Cabanac et al., 2006b], EGC'03 [Ghozzi et al., 2003b], EGC'01 [Ravat et al., 2001].

8.3 PERSPECTIVES

Nos propositions ont permis de définir une modélisation globale des données décisionnelles. Une des perspectives à court terme est d'offrir un langage d'analyse décisionnelle offrant un processus exploratoire des données multidimensionnelles brutes ou intégrant des

[13] Le contenu de chaque article est résumé dans le chapitre 6.

contraintes sémantiques et des versions d'étoile. Une réponse partielle à cette problématique est fournie dans le prochain chapitre.

A plus long terme, la personnalisation semble un terrain prometteur pour de nouvelles activités de recherche. L'aspect personnalisation peut intervenir au niveau des données comme nous l'avons proposé dans un premier temps, mais également au niveau des types de restitution. De plus, nous souhaitons poursuivre dans cette direction pour aboutir à de véritables systèmes adaptatifs qui tiennent compte des données mais également des actions réalisées pour proposer un véritable environnement de travail spécifique à chaque décideur.

CHAPITRE IV :

MANIPULATION DE

DONNEES

MULTIDIMENSIONNELLES

PLAN DU CHAPITRE

1 INTRODUCTION A L'ANALYSE MULTIDIMENSIONNELLE 83
1.1 Travaux existants 83
1.2 Problématique 83

2 ALGEBRE MULTIDIMENSIONNELLE 85
2.1 Constructeur 85
2.2 Noyau minimum fermé 86
2.2.1 Paramétrage d'une table multidimensionnelle 87
2.2.2 Présentation d'une table multidimensionnelle 87
2.2.3 Transformation d'une table multidimensionnelle 88
2.3 Opérateurs de second niveau 90
2.4 Opérateurs binaires 91
2.5 Adaptations aux schémas multidimensionnels étendus 93

3 LANGAGE GRAPHIQUE GOLAP 93
3.1 Visualisation d'un schéma multidimensionnel 94
3.2 Création initiale d'une table multidimensionnelle 95
3.3 Manipulations OLAP Graphiques 95
3.4 Complétude du langage graphique GOLAP 96

4 OLAP SQL 96
4.1 Langage de consultation d'OLAP SQL 97
4.1.1 La commande SELECT 97
4.1.2 Complétude du langage assertionnel 97
4.2 Langage de définition de OLAP-SQL 98
4.3 Langage de contrôle de OLAP-SQL 99
4.4 Langage de manipulation de OLAP-SQL 100

5 BILAN ET PERSPECTIVES 100
5.1 Bilan sur les langages de manipulation 100
5.2 Production 102
5.3 Perspectives 102

1 INTRODUCTION A L'ANALYSE MULTIDIMENSIONNELLE

Les modèles multidimensionnels reposent sur la métaphore du cube de données matérialisée au travers des faits et des dimensions. Les manipulations multidimensionnelles consistent à appliquer des processus exploratoires sur ces cubes afin de spécifier des cubes de données ou tableaux à *N* dimensions résultats [Tininini, 2003].

1.1 TRAVAUX EXISTANTS

Pour effectuer ces processus exploratoires, il est proposé différentes opérations de rotation, de forage, de sélection, de projection, d'ordonnancement [Li & Wang, 1996 ; Agrawal et al., 1997 ; Cabibbo & Torlonne, 1997 ; Gyssen & Lakshmanan, 1997 ; Cabibbo & Torlonne, 1998 ; Lehner 1998 ; Lehner et al., 1998 ; Marcel, 1998 ; Pedersen & Jensen, 1999 ; Mendelzon & Vaisman, 2000 ; Abello et al., 2003 ; Franconi & Kamble, 2004]. Ces langages de manipulation sont majoritairement définis à l'aide d'une algèbre [Agrawal et al., 1997 ; Cabibbo & Torlonne, 1997 ; Cabibbo & Torlonne, 1998 ; Gyssen & Lakshmanan, 1997 ; Lehner 1998 ; Lehner et al., 1998 ; Pedersen & Jensen, 1999 ; Abello et al., 2003 ; Franconi & Kamble, 2004 ; Ravat et al., 2006c). Ils peuvent être également définis à l'aide d'un langage déclaratif à base de d'un calcul [Li & Wang, 1996 ; Cabibbo & Torlonne, 1997] voire de règles [Marcel, 1998 ; Marcel, 1999 ; Mendelzon & Vaisman, 2000]. Ces opérations peuvent s'appliquer sur des tables relationnelles [Li & Wang, 1996], sur des classes d'objets [Lehner 1998 ; Mendelzon & Vaisman, 2000] mais plus majoritairement sur des cubes ou tableaux à *N* dimensions [Li & Wang, 1996 ; Agrawal et al., 1997 ; Cabibbo & Torlonne, 1997 ; Gyssen & Lakshmanan, 1997 ; Cabibbo & Torlonne, 1998 ; Marcel, 1998 ; Mendelzon & Vaisman, 2000 ; Abello et al., 2003 ; Franconi & Kamble, 2004].

Ces propositions présentent l'inconvénient de ne pas se soucier de la manière dont sont restituées les données aux décideurs. Comme indiqué dans le chapitre précédent, une visualisation sous forme de cubes en N dimensions (N>2) semble difficilement exploitable par les décideurs [Gyssen et al., 1997] et ce type de présentation occulte la hiérarchisation des dimensions. Ces propositions reposent sur des modèles de données logiques spécifiques et ne proposent pas toujours des définitions précises et formelles des différentes opérations. A l'heure actuelle, il n'existe pas de consensus sur la définition d'un noyau minimum complet offrant une algèbre d'interrogation multidimensionnelle. Enfin, seuls [Cabibbo & Torlonne, 1998] proposent la combinaison d'une algèbre et d'un langage graphique. Cependant le langage graphique se limite aux opérations de base, ce qui n'est pas complet au regard de l'algèbre.

En réponse à ces limites, nous souhaitons proposer une solution pour les langages de manipulation décisionnelle.

1.2 PROBLEMATIQUE

Le modèle de base et les extensions que nous avons présentés dans le chapitre précédent servent de support aux langages de manipulation OLAP qui font l'objet ce chapitre. Plus précisément, nous avons proposé le concept de Table Multidimensionnelle (TM) permettant de visualiser les données multidimensionnelles au travers de tableaux à double entrées hiérarchisées. Cette visualisation centre l'analyse sur un seul fait et facilite l'interprétation et l'analyse des données [Gyssens & Lakshmanan, 1997]. Ce concept servira de support à la définition des langages de manipulation de données décisionnelles. Suivant les utilisateurs et les objectifs assignés, nous souhaitons proposer différents langages de manipulation.

Dans un premier temps, nous avons défini une algèbre d'analyse décisionnelle. Comme pour toute algèbre, nous proposons un langage procédural permettant de combiner différentes opérations. Afin de répondre aux lacunes énoncées précédemment, nous souhaitons proposer une algèbre orientée décideur [Abelló et al., 2003]. Cette caractéristique induit :

- une définition d'opérateurs non pas sur des "structures simples" de niveau logique mais sur des composants conceptuels sémantiquement plus riches et plus proches des préoccupations des décideurs,
- une proposition d'un ensemble d'opérateurs exprimant les différentes actions que peut effectuer un utilisateur sur une structure de visualisation adaptée telle qu'une TM,
- la proposition d'une algèbre fermée présentant l'avantage de manipuler et de générer des TM et ainsi répondre à des requêtes complexes par la combinaison de différents opérateurs.

A l'instar de l'algèbre relationnelle qui offre un support complet et reconnu, nous souhaitons proposer un noyau minimal complet d'opérateurs unaires pouvant être combinés pour répondre aux besoins des décideurs. Ce noyau minimal sera complété par un ensemble d'opérateurs unaires de second niveau permettant de simplifier l'écriture des requêtes complexes. Enfin, cette algèbre proposera un ensemble d'opérateurs binaires de fusion de TM.

Afin d'être plus proche des préoccupations des décideurs, nous souhaitons également fournir un langage graphique de manipulation décisionnelle. Ce langage répond aux objectifs suivants :

- nous souhaitons faire abstraction de toute implantation physique des composants d'un schéma OLAP en proposant de travailler sur des représentations conceptuelles ;
- contrairement aux outils du marché présentant une simple vision arborescente des composants d'un schéma, nous souhaitons proposer une visualisation graphique explicite du schéma multidimensionnel tel que présenté dans le chapitre précédent ;
- contrairement aux outils du marché, l'expression de requêtes s'effectuera directement sur le graphe et de manière incrémentale ;
- ce langage graphique doit être complet au regard du noyau minimum de l'algèbre multidimensionnelle que nous souhaitons proposer ;
- toutes les actions graphiques peuvent être traduites en une suite de commandes SQL s'appliquant sur les structures de stockage R-OLAP.

Pour compléter cette proposition, nous souhaitons offrir un langage assertionnel spécifiquement adapté aux manipulations OLAP. Ce langage est plutôt destiné à un utilisateur averti afin de créer rapidement une TM analysable par les décideurs. Le langage assertionnel standard des bases relationnelles SQL n'est pas adapté et manque d'expressivité pour la définition d'une base multidimensionnelle. Aussi, nous proposons une extension de ce langage afin de conserver ses avantages (déclaratif, standard, puissant, reconnu) tout en facilitant la définition et la manipulation des concepts multidimensionnels :

- manipulation aisée des concepts multidimensionnels avec abstraction complète de l'implantation,
- définition standard des faits, dimensions et hiérarchies,
- consultation aisée avec une seule commande,
- indépendance par rapport à un SGBD physique.

Ces trois langages (algébrique, graphique et assertionnel) sont étudiés dans les sections suivantes.

2 ALGEBRE MULTIDIMENSIONNELLE

Notre algèbre propose un ensemble d'opérateurs permettant à un décideur de manipuler les composants d'une constellation afin d'effectuer ses analyses. Notre algèbre repose sur 4 types d'opérateurs : un opérateur de construction produisant une TM à partir d'une BDM, un noyau minimum fermé d'opérateurs fondamentaux portant sur les TM, un ensemble d'opérateurs avancés facilitant les manipulations OLAP en offrant des fonctionnalités de plus haut niveau et un ensemble d'opérateurs ensemblistes pour manipuler différents opérateurs.

2.1 CONSTRUCTEUR

Ce premier opérateur est spécifique car il permet de construire une première TM à partir des composants d'une BDM. Les autres opérateurs manipulent et génèrent des TM (propriété de fermeture).

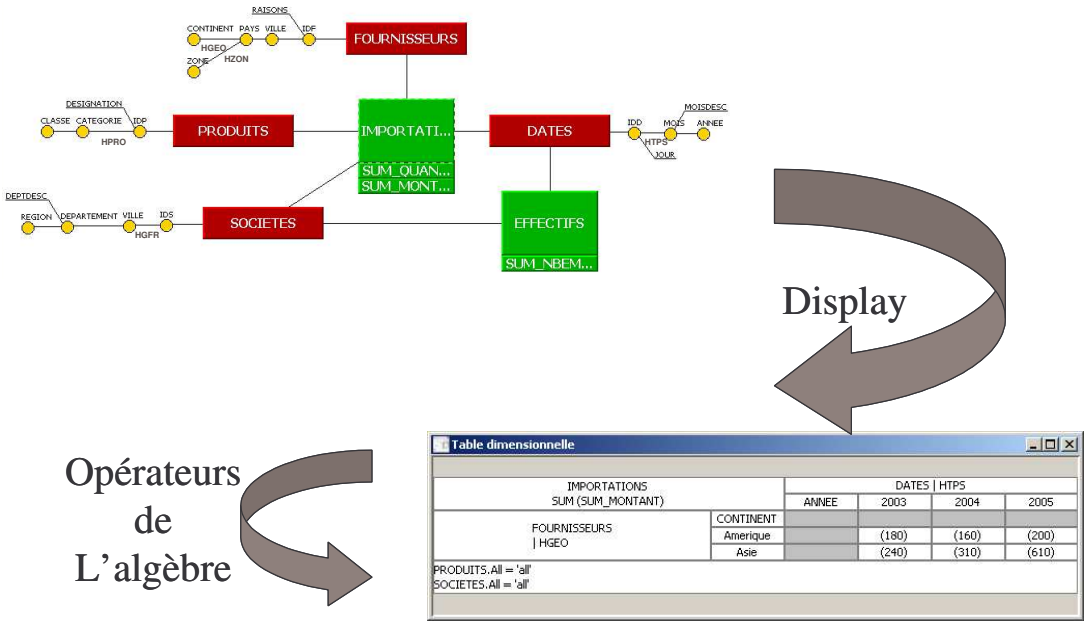


Figure 27 : Principes de l'algèbre multidimensionnelle

Définition. L'opération de construction est définie par

$$\mathbf{DISPLAY(N^{CS}, F, \{f_1(m_1), f_2(m_2),\ldots\}, DL, HL, DC, HC) = T_{RES}}$$

- N^{CS} est le nom de la constellation,
- F est le fait analysé (sujet de l'analyse),
- $\{f_1(m_1), f_2(m_2),\ldots\}$ est un ensemble de mesures $\{m_1, m_2,\ldots\}$ du fait F agrégées à l'aide de fonctions $f_1, f_2,\ldots$,
- DL est la dimension courante en ligne avec HL comme hiérarchie courante,
- DC est la dimension courante en colonne avec HC comme hiérarchie,
- $T_{RES}=(S_{RES}, L_{RES}, C_{RES}, R_{RES})$ est la TM résultat où S_{RES} =(F, $\{f_1(m_1), f_2(m_2),\ldots\}$), L_{RES}=(DL, HL, $<p^{DL}{}_1>$) et C_{RES}=(DC, HC, $<p^{DC}{}_1>$) paramètres de plus haute granularité de HL et HC et R_{RES}=true.

Exemple. Supposons que nous ayons le schéma en étoile suivant :

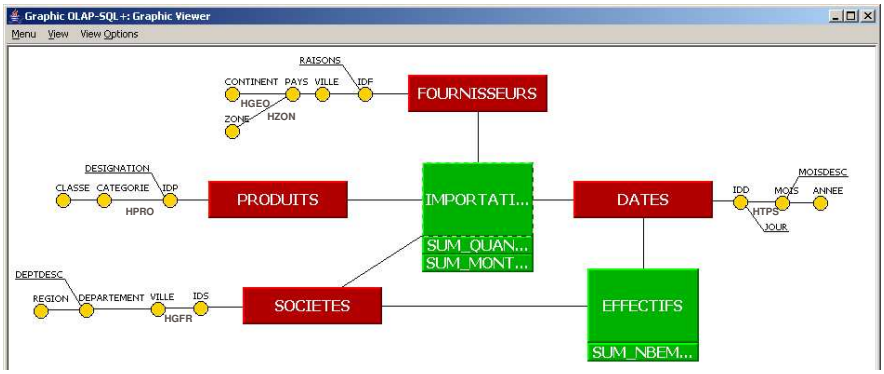


Figure 28 : Exemple d'une représentation graphique d'une constellation

Un décideur souhaite afficher la somme des montants importés par fournisseurs et par dates d'importations. La figure 2 présente le résultat de l'expression algébrique suivante : DISPLAY('SH_IMPORT', Importations, {*SUM*(Montant) }, Fournisseurs, HGeo, Dates, HTps) = T_{R1}

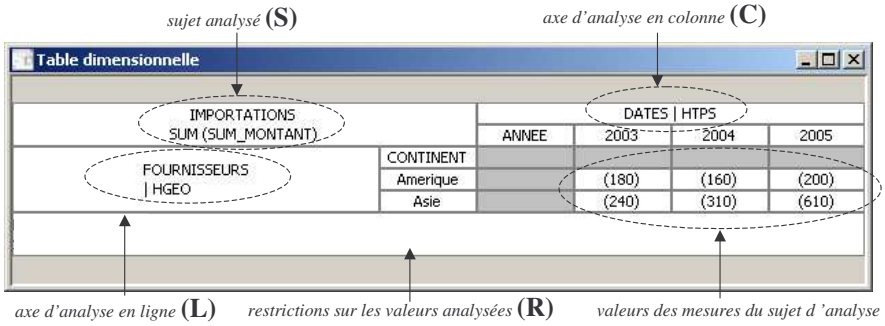


Figure 29 : Exemple d'une TM résultat après application de l'opérateur DISPLAY

2.2 NOYAU MINIMUM FERME

Ce noyau minimum permet de recenser l'ensemble des opérations de base qu'un décideur peut effectuer. Afin de faciliter la compréhension de ces opérateurs, nous proposons de les classifier en trois catégories : paramétrage, présentation et transformation d'une TM

En préambule, avant de définir ces différents opérateurs, nous proposons les notations suivantes. De part la propriété de fermeture, les opérateurs unaires prennent en entrée une TM source (notée T_{SRC}) et fournissent en sortie une autre TM notée T_{RES}. Chacune de ces deux TM est définie comme suit :

- $T_{SRC}=(S_{SRC}, L_{SRC}, C_{SRC}, R_{SRC})$ est la table multidimensionnelle initiale.
- $T_{RES}=(S_{RES}, L_{RES}, C_{RES}, R_{RES})$ est la table multidimensionnelle résultat. Par défaut, nous supposons que $S_{RES} = S_{SRC}$, $L_{RES} = L_{SRC}$, $C_{RES} = C_{SRC}$, $R_{RES}= R_{SRC}$; dans les définitions suivantes, nous spécifions uniquement les éléments de T_{RES} modifiés.

Dans les définitions suivantes **Att** représente soit un paramètre p, soit un paramètre avec une liste des attributs faibles $p(a^D_{p1}, a^D_{p2},\ldots)$, soit une liste d'attributs faibles $(a^D_{p1}, a^D_{p2},\ldots)$ d'un même paramètre p (non affiché dans le résultat).

2.2.1 Paramétrage d'une table multidimensionnelle

Les opérations de cette catégorie permettent de modifier les critères d'analyse multidimensionnelle en modifiant la structure d'affichage d'une table multidimensionnelle. Plus précisément, ces opérations permettent d'effectuer des rotations d'axes d'analyse, des forages vers le haut ou le bas afin de préciser le niveau de granularité d'analyse ou encore de sélectionner certaines données.

Le tableau ci-dessous décrit la syntaxe des opérations de cette catégorie.

Commande	Syntaxe
La rotation (DROTATE) permet, au sein d'une TM, soit de changer un axe d'analyse par un autre, soit de changer la hiérarchie sur un même axe	**DROTATE(T_{SRC}, D_{old}, D_{new} ,H^{Dnew}_{k}) = T_{RES}** – D_{old} est la dimension sur laquelle s'applique la rotation, $D_{old} \in$ {DC, DL}, – D_{new} est la nouvelle dimension qui remplace D_{old} dans T_{RES}, – H^{Dnew}_{k} est la hiérarchie courante utilisée pour graduer la dimension D_{new} (le paramètre de granularité maximale est choisi), – T_{RES} est la TM résultat où seuls sont modifiés les éléments suivants : o Si D_{old}=DL alors L_{RES}=(D_{new}, H^{Dnew}_{k}, <p^{DL}_{1}>), o Si D_{old}=DC alors C_{ES}=(D_{new}, H^{Dnew}_{k}, <p^{DC}_{1}>).
Les forages vers le bas ou vers le haut (DRILL-DOWN ou ROLL-UP) permettent au décideur d'analyser les données de manières plus ou moins détaillées en modifiant les différents niveaux de graduation utilisés pour visualiser les données. Un niveau de graduation est représenté soit par un paramètre soit par ses attributs faibles soit les deux composants ensemble.	**DRILLDOWN(T_{SRC}, D, Att_{inf}) = T_{RES}** – D est la dimension sur laquelle s'opère le forage, – Att_{inf} représente un attribut inférieur dans la hiérarchie courante. Les niveaux de graduation intermédiaires entre la graduation inférieure de la table initiale et la nouvelle graduation ne sont pas pris en compte dans la table résultat. – T_{RES} est la TM résultat où seuls sont modifiés les éléments suivants : o Si D=DL alors L_{RES}=(D, H^{D}_{k}, <p^{DL}_{1},…, p^{DL}_{v}, Att_{inf}>), o Si D=DC alors C_{RES}=(D, H^{D}_{k}, <p^{DC}_{1},… p^{DC}_{w}, Att_{inf}>).
	ROLLUP(T_{SRC}, D, Att_{sup}) = T_{RES} – D est la dimension sur laquelle s'opère le forage, – Att_{sup} représente le niveau de graduation supérieur utilisé dans la table résultat, les graduations inférieures présentes dans la table initiale sont supprimées de la table résultat. – T_{RES} est la TM résultat où seuls sont modifiés les éléments suivants : o Si D=DL alors L_{RES}=(D, H^{D}_{k}, <p^{DL}_{1},…, Att_{sup}>), o Si D=DC alors C_{RES}=(D, H^{D}_{k}, <p^{DC}_{1},…, Att_{sup}>).
La sélection (SELECT) permet de restreindre l'ensemble des valeurs affichées. Ces restrictions portent aussi bien sur les valeurs des attributs des dimensions que celles des mesures du fait.	**SELECT(T_{SRC}, pred) = T_{RES}** – pred est un prédicat normalisé (conjonction de disjonctions) de sélection portant sur les dimensions et/ou le fait, – T_{RES} est la TM résultat où seul est remplacé le prédicat R_{RES}=pred.

Figure 30 : opérateurs de paramétrage d'une table multidimensionnelle

2.2.2 Présentation d'une table multidimensionnelle

Les opérations de cette catégorie permettent de préciser la structure de visualisation des données (classement des valeurs de paramètres voire des paramètres eux-mêmes, ainsi que ajout et suppression de fonctions d'agrégation).

Le tableau ci-dessous décrit la syntaxe des opérations de cette catégorie.

Commande	Syntaxe
L'opération de classement (SWITCH) intervertit deux valeurs d'un attribut d'une dimension pour permettre l'ordonnancement des valeurs affichées	**SWITCH(T_{SRC}, D, Att, v_1, v_2) = T_{RES}** – T_{SRC} est laTM initiale où dom(Att)=<…v_1,…v_2,…>, – D est la dimension contenant le paramètre sur lequel s'applique permutation, – Att est l'attribut sur lequel la permutation s'effectue, – v_1 et v_2 sont les valeurs permutées, – T_{RES} est la TM résultat où dom(Att)=<…v_2,…v_1,…>.
L'imbrication (NEST) permet d'intégrer, dans les dimensions d'une TM, les données provenant d'une ou plusieurs dimensions. Elle permet d'utiliser les paramètres de plusieurs dimensions dans l'espace 2D de la TM	**NEST(T_{SRC}, D, Att, D_{nested}, Att_{nested}) = T_{RES}** – D est la dimension sur laquelle s'opère l'imbrication, – Att est l'attribut au niveau duquel l'imbrication est effectuée, – D_{nested} est la dimension d'où est issu l'attribut imbriqué, – Att_{nested} est le paramètre et/ou les attributs faibles de D_{nested} imbriqués. – T_{RES} est la TM résultat où seuls sont modifiés les éléments suivants : o Si D=DL alors L_{RES}=(D, H^D_k, <p^{DL}_1,…, Att, Att_{nested}>), o Si D=DC alors C_{RES}=(D, H^D_k, <p^{DC}_1,…, Att, Att_{nested}>).
L'opération de calculs d'agrégats (AGREGATE) permet d'ajouter dans une TM des calculs agrégeant les lignes et/ou les colonnes. Cette opération correspond à l'opération Cube proposée par (Gray, *et al.* 1996). L'opération UNAGREGATE permet de supprimer l'affichage des valeurs affichées.	**AGREGATE(T_{SRC}, D, F(Att)) = T_{RES}** – D est la dimension sur laquelle s'applique l'agrégation, – F représente une fonction d'agrégation (sum, avg, count…) appliquée sur un paramètre et/ou un attribut faible noté Att, pour lequel dom(Att)=<v_1,…,v_x>. – T_{RES} est la TM résultat où ∀i∈ [1..x], dom(Att) = <v_1, $F(v_1)$,…,v_x, $F(v_x)$>. Chaque valeur initiale est décomposée en deux valeurs, la valeur elle-même et une valeur représentant l'agrégation de celle-ci.
	UNAGREGATE(T_{SRC}) = T_{RES} – T_{RES} est la TM résultat où toutes les valeurs agrégées sont éliminées.

Figure 31 : opérateurs de présentation d'une table multidimensionnelle

2.2.3 Transformation d'une table multidimensionnelle

Les opérations de cette catégorie permettent soit de modifier un fait par ajout ou suppression de mesures, soit de modifier une dimension par transformation d'une mesure en paramètre ou vice versa.

Le tableau ci-dessous décrit la syntaxe des opérations de cette catégorie.

Commande	Syntaxe
Les opérations d'ajout (ADDM) et de suppression (DELM) de mesures permettent de modifier l'ensemble des mesures affichées sur la table multidimensionnelle résultat.	**ADDM(T_{SRC}, $f_i(m_i)$) = T_{RES}** – $f_i(m_i) \notin \{f_1(m_1),\ldots,f_x(m_x)\}$ dans T_{SRC} est une mesure qui doit être ajoutée au fait courant dans la TM, – T_{RES} est la TM résultat où $S_{RES} = (F_{SRC}, \{f_1(m_1),\ldots,f_x(m_x), f_i(m_i)\})$.
	DELM(T_{SRC}, $f_i(m_i)$) = T_{RES} – $f_i(m_i) \in \{f_1(m_1),\ldots,f_i(m_i),\ldots,f_x(m_x)\}$ dans T_{SRC} est une mesure qui doit être supprimée du fait courant dans la TM telle que $S_{RES} = (F_{SRC}, \{f_1(m_1),\ldots,f_i(m_i),\ldots,f_x(m_x)\})$. – T_{RES} est la TM résultat où seul est modifié $S_{RES} = (F_{SRC}, \{ f_1(m_1),\ldots,f_{i-1}(m_{i-1}), f_{i+1}(m_{i+1}),\ldots,f_x(m_x)\})$.
L'opération PUSH transforme un paramètre afin qu'il apparaisse dans la TM à l'utilisateur comme une mesure.	**PUSH(T_{SRC}, D, Att) = T_{RES}** – D est la dimension contenant l'attribut de conversion $Att \in H^D_k$ où H^D_k est la hiérarchie courante, – Att est le paramètre ou l'attribut faible convertit en mesure. – T_{RES} est la TM résultat où $S_{RES} = (F_{SRC}, \{f_1(m_1), f_2(m_2),\ldots, Att\})$ avec $Att \notin H^D_k$.
L'opération de conversion PULL transforme une mesure en paramètre de la TM. Les valeurs de la mesure sont affichées au niveau des entêtes de ligne ou de colonne.	**PULL(T_{SRC}, $f_i(m_i)$, D) = T_{RES}** – $f_i(m_i)$ est une mesure du fait courant affiché dans la TM initiale, – D est la dimension (ligne ou colonne) dans laquelle la mesure est transférée, – T_{RES} est la TM résultat où seuls sont modifiés les éléments suivants : o Si D=DL alors $L_{RES}=(D, H^D_k, <p^{DL}_1,\ldots,f_i(m_i)>)$, o Si D=DC alors $C_{RES}=(D, H^D_k, <p^{DC}_1,\ldots,f_i(m_i)>)$.

Figure 32 : opérateurs de transformation d'une table multidimensionnelle

Exemple. Le décideur poursuit l'analyse précédente en focalisant son observation sur le montant et la moyenne des montants des importations en 2005 de produits électroniques. Il souhaite également affiner l'analyse en visualisant les montants plus finement, par pays d'origines de chaque fournisseur, tout en modifiant l'axe des colonnes pour observer les mesures par société importatrice. Pour ce faire, cette requête complexe est spécifiée par combinaison de plusieurs opérateurs élémentaires du noyau de l'algèbre :

- une sélection des PRODUITS 'Electronique' et des DATES en 2005,
- une opération de forage vers le bas sur l'axe des FOURNISSEURS,
- une opération de rotation des dimensions DATES et SOCIETES,
- une opération d'ajout de la mesure AVG(Montant).

L'expression algébrique suivante produit la TM décrite en figure 3.

DROTATE(ADDM(SELECT(DRILLDOWN(T_{R1}, Fournisseurs, Pays), Produits.Classe = 'Electronique' ∧ Dates.Annee = 2005), AVG(Montant)) , Fournisseurs, Societes, HGFr) = T_{R2}.

Table dimensionnelle

IMPORTATIONS SUM (SUM_MONTANT), AVG (SUM_MONTANT)			SOCIETES | HGFR	
			REGION	Midi-Pyrenees
FOURNISSEURS | HGEO	CONTIN...	PAYS		
	Amerique	Bresil		(100, 100)
		Etats-Unis		(100, 100)
	Asie	Chine		(230, 230)
		Inde		(160, 160)
		Thailande		(220, 220)
PRODUITS.CLASSE = 'Electronique' DATES.ANNEE = 2005				

Figure 33 : TM résultat de la combinaison des opérateurs

2.3 OPERATEURS DE SECOND NIVEAU

Le noyau minimum de l'algèbre offre la possibilité de visualiser et d'effectuer des analyses plus ou moins complexes sur les données d'une constellation. Cependant, certaines analyses complexes nécessitent de nombreuses combinaisons d'opérateurs élémentaires du noyau. Afin d'améliorer le traitement des requêtes complexes, nous proposons un ensemble d'opérateurs de second niveau (construits par combinaison d'opérateurs du noyau minimum). L'intérêt de cette proposition est double : l'expression des analyses est réduite et les traitements systèmes correspondants aux opérations de second niveau peuvent être optimisés par rapport à la combinaison équivalente d'opérateurs du noyau.

Les opérateurs de second niveau peuvent intervenir dans différents domaines. Au niveau de la rotation, nous proposons les deux opérateurs suivants :

- l'opération de rotation de hiérarchies (HROTATE) consiste simplement à changer la hiérarchie courante d'une dimension ligne ou colonne ;
- l'opération de rotation de faits (FROTATE), équivalente à l'opération Drill-Accross proposée par [Abelló et al., 2003], consiste à utiliser un nouveau fait dans la TM tout en conservant les caractéristiques des axes d'analyse courants. Cette opération n'est applicable que lorsque le nouveau fait partage au moins les deux dimensions courantes du fait de la TM initiale.

Pour les forages, l'opération de projection d'un paramètre (PLOT) que nous proposons consiste à afficher les données suivant un paramètre quelconque de la dimension. Pour l'ordonnancement (croissant ou décroissant) des valeurs d'un attribut de dimension, nous proposons l'opération ORDER. Au niveau de la sélection, l'opération de "désélection" (UNSELECT) consiste à annuler toutes les sélections sur les dimensions et le fait.

Le tableau suivant présente les différents opérateurs de second niveau et l'expression représentant la combinaison équivalente d'opérateurs du noyau. Au préalable, nous présentons deux composants essentiels pour la compréhension de ces transformations.

- Toute dimension possède un paramètre système nommé "All". Ce paramètre est le paramètre de plus haute granularité de la dimension auquel sont reliés les paramètres de plus haute granularité définis par le concepteur. Il est composé d'une seule valeur "all" et regroupe toutes les instances de la dimension ;
- History(T_{old}, obj, T_{new})=T_R est une fonction matérialisant l'historique des opérations qui ont été appliquées dans T_{old} sur obj (fait ou dimension) et qui doivent s'appliquer à T_{new}.

Opérateur	Combinaison équivalente d'opérateurs du noyau
HROTATE(T_{SRC}, D, H^{D}_{k})=T_{RES}	DROTATE(T_{SRC}, D, D ,H^{D}_{k})=T_{RES}
PLOT(T_{SRC}, D, Niv)=T_{RES}	DRILLDOWN(ROLLUP(T_{SRC}, D, All), D, Niv)=T_{RES}
ORDER(T_{SRC}, D, p, ord)=T_{RES} ord∈ {'asc', 'dsc'}	SWITCH(…(SWITCH(T_{SRC}, D, p, v_1, v_2), …), D, p, v_3, v_4)=T_{RES}
FROTATE(T_{SRC}, F_{new}, {$f_1(m_1)$, $f_2(m_2)$,…})=T_{RES}	History(T_{SRC}, DL, History (T_{SRC}, DC, DISPLAY(N^{CS}, F_{new}, {$f_1(m_1)$, $f_2(m_2)$,…}, DL, HL, DC, HC)))=T_{RES}
UNSELECT(T_{SRC})=T_{RES}	SELECT(T_{SRC}, F_{SRC}.All='all' ∧ D_{SRC_1}.All='all' ∧ … ∧ D_{SRC_q}.All='all')=T_{RES}

Figure 34 : opérateurs de second niveau

Exemple. Le décideur modifie l'analyse de l'exemple précédent en focalisant son observation sur la somme et la moyenne des montants des importations de 2005 de produits électroniques. Il souhaite également affiner l'analyse en visualisant les montants plus finement, par pays d'origine de chaque fournisseur tout en modifiant l'axe des colonnes pour observer les mesures par sociétés importatrices. En réponse à ce besoin, l'expression algébrique suivante produit la TM décrite en figure 5.

PLOT(HROTATE(UNSELECT(DELM(T_{R2}, AVG(Montant))), Fournisseurs, HZon), Societes, Ville) = T_{R3}.

Table dimensionnelle

IMPORTATIONS SUM (SUM_MONTANT)		SOCIETES | HGFR			
		VILLE	Bordeaux	Lyon	Toulouse
FOURNISSEURS | HZON	ZONE				
	E		(140)	(100)	(1160)
	O		(180)	(200)	(540)

Figure 35 : TM résultat de la combinaison des opérateurs de second niveau

2.4 OPERATEURS BINAIRES

Les décideurs sont couramment confrontés à un problème peu étudié [Ravat et al., 2002] [Benitez-Guerrero et al., 2003], à savoir la fusion du contenu de tables dimensionnelles. Or, la fusion de deux TD répond à un besoin de corrélation nécessaire lors de la prise de décision. Par exemple, supposons qu'un décideur possède une table dimensionnelle TD1 contenant les ventes des produits pour 2002 et une table TD2 contenant les ventes des produits pour 2003. Une première fusion entre ces deux tables lui permettrait de calculer le montant total des ventes de produits en 2002 et 2003. Une seconde fusion pourrait calculer la différence des ventes et donc mettre rapidement en avant la variation des parts de marché. Ces calculs sont actuellement réalisés de manière empirique, et souvent fastidieuse, en effectuant des extractions et des recopies dans des tableurs. Une première solution [Franconni & Kamble, 2004] permet d'effectuer l'union, l'intersection et la différence de deux cubes. Cette solution se limite à deux cubes ayant une structure strictement identique et les arêtes de ces cubes ne reposent pas sur des axes hiérarchisés.

Cette section vise à étendre notre algèbre afin de compléter les opérateurs unaires présentés dans les sections précédentes par un opérateur binaire manipulant deux TM en entrée pour en

générer une seule en sortie. Cet opérateur de fusion a pour vocation de favoriser la combinaison de plusieurs TD et donc faciliter les corrélations d'analyses.

Pour que l'opération de fusion puisse s'appliquer, il faut deux tables compatibles. Soient deux tables $T_{SRC1}=(S_{SRC1}, L_{SRC1}, C_{SRC1}, R_{SRC1})$ et $T_{SRC2}=(S_{SRC2}, L_{SRC2}, C_{SRC2}, R_{SRC2})$ telles que $\forall i \in [1..2]$,

- $S_{SRCi} = (F, \{m^{SRCi}_1,\ldots, m^{SRCi}_s\})$,
- $L_{SRCi} = (DL^{SRCi}, HL^{SRCi}, <p^{DL/SRCi}_1,\ldots, p^{DL/SRCi}_{cl}>)$,
- $C_{SRCi} = (DC^{SRCi}, HC^{SRCi}, <p^{DC/SRCi}_1,\ldots, p^{DC/SRCi}_{cc}>)$,
- $R^{SRCi} = pred^{SRCi}_1 \wedge\ldots\wedge pred^{SRCi}_t$.

Définition. Deux tables T_{SRC1} and T_{SRC2} sont compatibles si et seulement si :

- S_{SRC1} and S_{SRC2} sont compatibles : ils possèdent le même nombre de mesures notées $\{m^{SRC1}_1,\ldots, m^{SRC1}_s\}$ et $\{m^{SRC2}_1,\ldots, m^{SRC2}_s\}$, et $\forall i \in [1..s]$, les mesures m^{SRC1}_i et m^{SRC2}_i sont compatibles deux à deux ; autrement dit, le domaine de définition de ces mesures est le même $\forall i \in [1..s]$, $dom(m^{SRC1}_i) = dom(m^{SRC2}_i)$,
- L_{SRC1} and L_{SRC2} sont compatibles : ils ont la même structure (même dimension $DL^{SRC1}=DL^{SRC2}$, même hiérarchie $HL^{SRC1}=HL^{SRC2}$, même ensemble ordonné de paramètres affichés dans les dimensions $<p^{DL\ SRC1}_1,\ldots, p^{DL\ SRC1}_{cl}>=<p^{DL\ SRC2}_1,\ldots, p^{DL\ SRC2}_{cl}>$). Notons que les valeurs des attributs des dimensions ne sont pas nécessairement identiques.
- C_{SRC1} and C_{SRC2} sont compatibles : ils possèdent la même structure.

Nous avons également défini la propriété de semi-compatibilté permettant aux opérations binaires de s'appliquer sur des tables non strictement compatibles.

Definition. Deux tables T_{SRC1} and T_{SRC2} sont semi-compatibles si et seulement si :

- L_{SRC1} and L_{SRC2} sont compatibles,
- C_{SRC1} and C_{SRC2} sont compatibles.

L'opérateur SET permettant de manipuler deux tables compatibles ou semi-compatibles est défini comme suit :

Definition. $SET(T_{SRC1}, T_{SRC2} [, Calc]) = T_{RES}$ où

$SET \in \{UNION, INTERSECT, MINUS\}$, T_{SRC1} et T_{SRC2} sont deux tables compatibles et T_{RES} est la TM résultat. "Calc" est une fonction de calcul qui est appliquée aux mesures correspondantes dans les deux tables T_{SRC1} et T_{SRC2}. Il faut noter que la fonction de calcul "Calc" est un paramètre optionnel et s'il n'est pas spécifié, les mesures des tables T_{SRC1} et T_{SRC2} ne seront pas regroupées dans la table T_{RES}.

- $S_{RES} = (F^{SRC1}\_F^{SRC2}, \{m^{SRC1}_1,\ldots, m^{SRC1}_s\})$ si *Calc* est spécifié,
- $S_{RES} = (F^{SRC1}\_F^{SRC2}, \{m^{SRC1}_1,\ldots, m^{SRC1}_{s1}, m^{SRC2}_1,\ldots, m^{SRC2}_{s2}\})$ si *Calc* n'est pas spécifié,
- $L_{RES} = (DL^{SRC1}, HL^{SRC1}, <All, p^{DL\ RES}_1,\ldots, p^{DL\ RES}_{cl}>)$ où $\forall i \in [1..cl]$,

  Si SET = UNION, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \cup dom(p^{DL\ SRC2}_i)$,

Si SET = INTERSECT, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \cap dom(p^{DL\ SRC2}_i)$,

Si SET = MINUS, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \setminus dom(p^{DL\ SRC2}_i)$,

- $C_{RES} = (DC^{SRC1}, HC^{SRC1}, <All, p^{DC\ RES}_1, \ldots, p^{DC\ RES}_{cc}>)$ où $\forall i \in [1..cc]$,

  Si SET = UNION, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \cup dom(p^{DL\ SRC2}_i)$,

  Si SET = INTERSECT, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \cap dom(p^{DL\ SRC2}_i)$,

  Si SET = MINUS, $dom(p^{DL\ RES}_i) = dom(p^{DL\ SRC1}_i) \setminus dom(p^{DL\ SRC2}_i)$,

- Si SET = UNION, $R_{RES} = R_{SRC1} \vee R_{SRC2}$, Si SET = INTERSECT, $R_{RES} = R_{SRC1} \wedge R_{SRC2}$, Si SET = INTERSECT, $R_{RES} = R_{SRC1} \wedge \neg R_{SRC2}$.

2.5 ADAPTATIONS AUX SCHEMAS MULTIDIMENSIONNELS ETENDUS

Les différentes extensions du modèle de base nécessitent le développement de compléments pour les opérateurs algébriques. Dans un premier temps, nous avons apporté des solutions dans deux cas :

- manipulation de données contraintes,
- manipulation de données personnalisées.

Pour la manipulation de données contraintes, nous avons orienté nos recherches vers la proposition d'un langage d'interrogation évitant les corrélations incohérentes et précisant l'ensemble des données à visualiser. Vous trouverez tous les détails de ce langage dans la thèse de Faiza Ghozzi [Ghozzi, 2004b]. Notamment, ces extensions portent sur les opérations de forage : lors du passage d'un paramètre spécifique à hiérarchie vers un paramètre commun à plusieurs hiérarchies, le décideur précise la portée des données à analyser à l'aide d'un booléen (maintien des valeurs spécifiques ou nouvelle analyse comprenant l'ensemble des données). De même en fonction des contraintes, les rotations sont autorisées ou pas ; dans le cas où elles sont autorisées, il faut préciser la portée de l'analyse (maintien des valeurs spécifiques ou nouvelle analyse).

La personnalisation proposée consiste à associer aux attributs des poids reflétant l'intérêt prioritaire que porte l'utilisateur aux attributs d'une constellation. Ces priorités sont exploitées par le système lors des manipulations utilisateurs. Les travaux d'extensions consistent à intégrer un seuil optionnel aux opérations de construction (DISPLAY), de forages (DRILLDOWN, ROLLUP) et de rotation (DROTATE). Par exemple, la table multidimensionnelle résultat d'une opération DISPLAY ou DROTATE n'affiche pas nécessairement le paramètre de plus haute granularité des dimensions concernées (fonctionnement classique) mais contient automatiquement tous les éléments dont le poids est supérieur au seuil. Cette version étendue présente l'avantage d'offrir un affichage personnalisé tout en diminuant le nombre de commandes à exécuter (notamment les forages vers le bas). Lors des forages vers le bas, le système affiche automatiquement le paramètre demandé et tous les paramètres intermédiaires dont le poids est supérieur au seuil. Cette solution présente l'avantage de limiter le nombre d'actions à réaliser.

3 LANGAGE GRAPHIQUE GOLAP

Pour compléter ce langage algébrique, nous proposons un langage graphique. Ce langage présente l'avantage de présenter les données du schéma multidimensionnel sous une forme graphique. Ce schéma graphique sert de support à l'élaboration incrémentale et graphique de requêtes. Dans les sections suivantes, nous étudions ces principes d'affichage des données et de requêtage graphique. Pour terminer, nous étudions la complétude de ce langage par rapport à notre algèbre.

3.1 VISUALISATION D'UN SCHEMA MULTIDIMENSIONNEL

Notre objectif est d'offrir une vue globale des données analysables. Nous représentons une constellation par un graphe où chaque nœud est un fait ou une dimension et chaque arc représente un lien entre un fait et une dimension. Les nœuds se différencient grâce à des couleurs différentes : vert pour un fait et rouge pour une dimension. Cette représentation peut être développée en déployant les hiérarchies constituant chaque dimension. La visualisation développée du graphe introduit alors d'autres types de nœuds représentant les paramètres et les attributs faibles. Chaque dimension est alors constituée par un sous graphe de paramètres et d'attributs faibles hiérarchiquement liés.

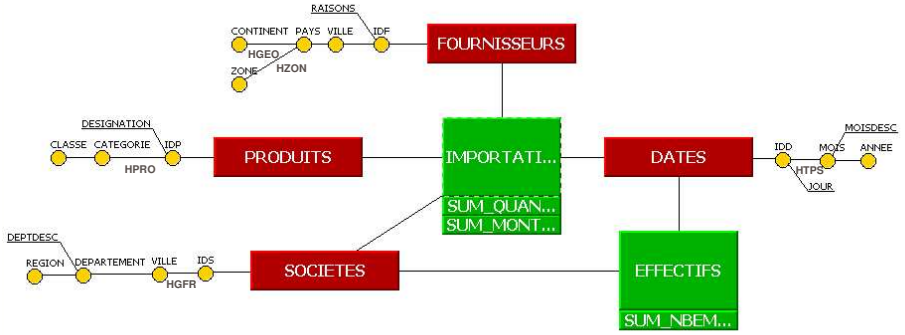


Figure 36 : Représentation développée d'un schéma conceptuel

Pour faciliter la visualisation des données, nous proposons d'afficher le sous-arbre d'une dimension (l'ensemble de ces hiérarchies) de deux manières. La première, la vision compacte, permet d'afficher un seul arbre pour l'ensemble des paramètres et attributs faibles ; chaque chemin de la racine à une feuille constitue une hiérarchie. La seconde, la vision éclatée, permet de construire un arbre pour chaque hiérarchie et affichant en double les attributs de la dimension partagée par deux hiérarchies.

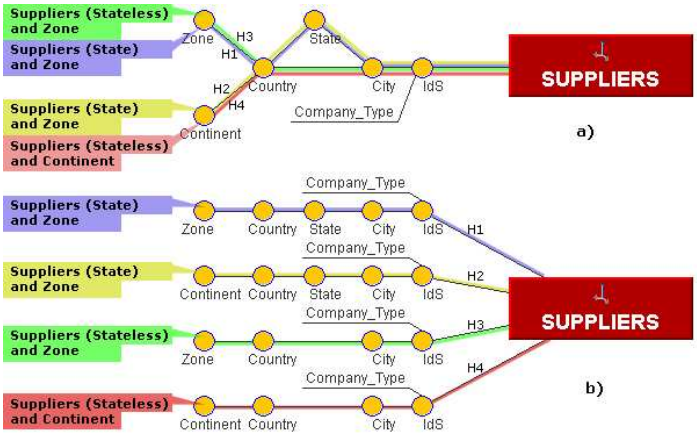


Figure 37 : Vision compacte (a) ou éclatée (b) d'une dimension

Dans le cas d'un schéma composé d'un grand nombre de composants, nous avons proposé une vision basée sur un arbre hyperbolique représenté dans un espace euclidien. L'utilisateur a une vision du schéma sous la forme d'un graphe sur lequel il peut faire des rotations pour se centrer sur la partie du graphe qu'il souhaite analyser

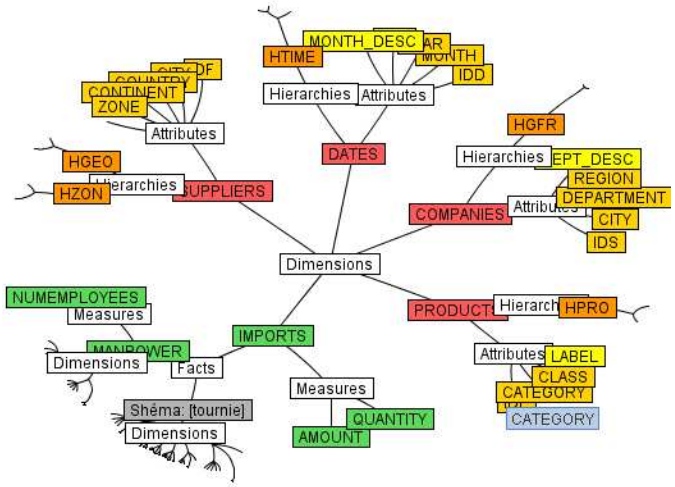


Figure 38 : Vision hyperbolique d'une constellation

3.2 CREATION INITIALE D'UNE TABLE MULTIDIMENSIONNELLE

La construction d'une TM s'effectue par manipulation directe et incrémentale du graphe. L'utilisateur sélectionne trois nœuds devant être un fait et deux dimensions ; le système assure la cohérence des sélections en rendant inaccessibles les nœuds invalides au fur et à mesure des sélections opérées par l'utilisateur. Ces différentes manipulations graphiques correspondent à l'opération DISPLAY de l'algèbre présentée dans la section précédente.

Exemple. La TM de la figure 5 a été obtenue par manipulations graphiques et utilisation de menus contextuels. Le décideur a sélectionné le fait IMPORTATION, puis au travers d'un formulaire il a spécifié les mesures et les fonctions d'agrégation utilisées (❶). Ensuite, il a sélectionné incrémentalement les dimensions FOURNISSEURS (❸) et DATES (❷) ; à chaque sélection un formulaire a permis de préciser la hiérarchie, les paramètres et/ou les attributs faibles affichés. Le décideur a validé la requête et le système produit en résultat une TM (❹).

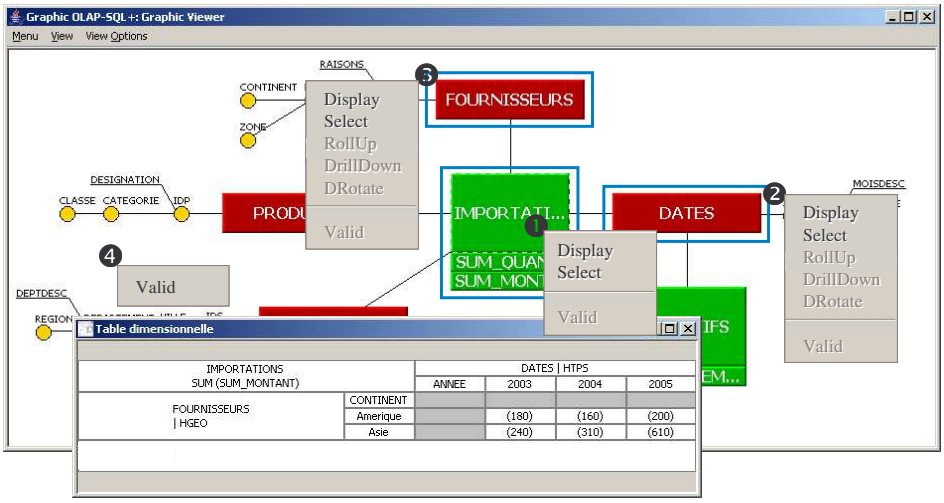


Figure 39 : Construction graphique d'une table multidimensionnelle

3.3 MANIPULATIONS OLAP GRAPHIQUES

Le système offre à un décideur deux manières d'effectuer des analyses décisionnelles. Un décideur peut agir graphiquement sur la constellation, mais il peut également appliquer certaines opérations directement sur les TM.

Exemple. Après construction de la TM précédente, le décideur poursuit son analyse. Plus précisément, (❶) il affine son analyse en affichant les fournisseurs par pays d'origine, et (❷) il focalise son analyse en sélectionnant la catégorie de produit intitulée 'Electronique'. Le forage de la dimension FOURNISSEURS (❶) est exprimé soit sur le graphe, soit directement sur la TM obtenue précédemment. La sélection permettant de focaliser l'analyse uniquement sur les produits électroniques s'exprime en manipulant la dimension PRODUITS sur le graphe (❷). Cette opération ne peut être définie qu'à partir du graphe puisque la dimension impliquée n'est pas disponible dans la TM.

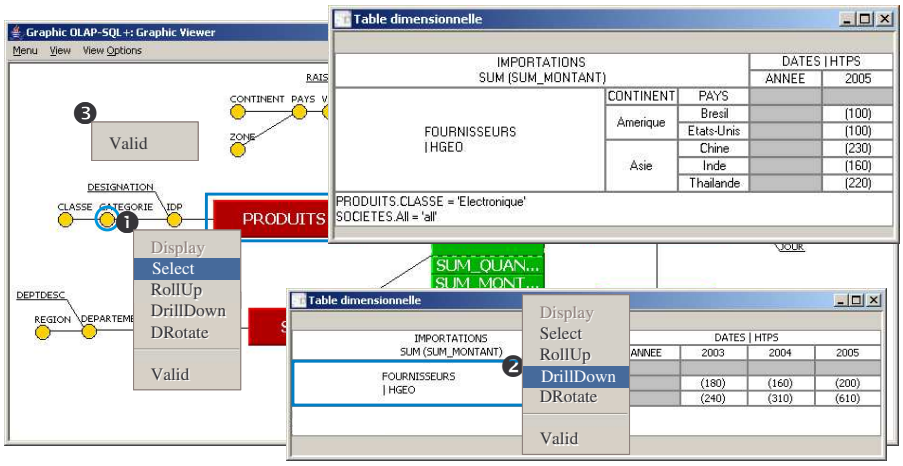


Figure 40 : Construction d'une requête.

3.4 COMPLETUDE DU LANGAGE GRAPHIQUE GOLAP

Toute opération de l'algèbre est exprimable au travers du langage GOLAP qui est donc complet au regard de notre algèbre multidimensionnelle.

		Opérateurs du noyau												Opérateurs avancés				
		Display	Drotate	DrillDown	RollUp	Nest	Switch	Aggregate	Unaggregate	Push	Pull	Addm	Delm	Select	Hrotate	Plot	Order	Frotate	Unselect
Glisser Déposer	Schéma -> TM	x	x	x	x	x				x	x		x	x	x			x	
	TM -> TM					x	x			x	x	x	x	x					x
Menu Contextuel	Schéma -> TM	x	x	x	x	x		x		x	x	x	x	x	x	x	x	x	
	TM -> TM			x	x		x	x	x			x	x	x		x	x		x
Opérations supportées par GOLAP		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Le tableau ci-dessus permet de préciser comment s'effectue chacune des opérations de l'algèbre à l'aide des commandes graphiques de GOLAP.

Ces commandes correspondent à :

- soit des "Glisser/Déposer" du schéma vers une TM ou d'une TM vers une autre,
- soit l'utilisation de menus contextuels.

4 OLAP SQL

Cette section présente nos travaux relatifs au développement d'un langage assertionnel spécifiquement adapté aux schémas de données OLAP. SQL étant le langage assertionnel standard le plus connu, nous proposons de l'étendre pour définir, consulter, manipuler ou contrôler une base de données multidimensionnelles. Dans une première section, nous présentons la commande de consultation des données décisionnelles et sa complétude par rapport à l'algèbre définie en section 2. Dans les sections suivantes, nous présentons les fonctionnalités des langages de définition, de manipulation et de contrôle.

4.1 LANGAGE DE CONSULTATION D'OLAP SQL

Notre objectif est de proposer un langage d'interrogation uniforme d'une base de données en constellation. A l'instar du langage SQL qui permet d'exprimer simplement l'ensemble des opérations de l'algèbre relationnelle, notre langage OLAP-SQL a pour objectif d'exprimer l'ensemble des opérations de l'algèbre multidimensionnelle (Ravat et al., 2001) (pour une base en constellation). Notre commande de consultation présente l'avantage d'être plus complète que les propositions des SGBD commerciaux (extension du GROUP BY avec Oracle, extension du SELECT avec MS SQL Server).

4.1.1 La commande SELECT

Nous proposons une commande unique permettant l'expression de l'ensemble des opérations multidimensionnelles. Cette commande repose sur l'extension de la requête de recherche du langage SQL (select).

select <agreg>(<mesure1>), <agreg>(<mesure2>), …
[**according to** [**rows** <parametre1_lig>, <parametre2_lig>,…,
of <nomdimension_lig> **with** <nomhierarchie_lig>,]
[**columns** <parametre1_col>, <parametre2_col>,…,]
of <nomdimension_col> **with** <nomhierarchie_col>]]
from <nomfait>
[**where** <predicat>]
[**order by** <parametre1> **values**(<valeur1_1>, <valeur1_2>,…)
[…[,<parametre2> **values**(<valeur2_1>, <valeur2_2>,…)],…]];

Cette commande de consultation regroupe différentes clauses :

- <**select**> définit des mesures à afficher ;
- <**according to**> définit des paramètres en fonction desquels les mesures sont calculées. Cette clause permet de structurer le résultat obtenu en fonction de lignes et de colonnes correspondant aux dimensions ;
- <**from**> précise le fait de l'analyse ;
- <**where**> est une clause optionnelle permettant de restreindre le résultat obtenu. La restriction peut s'appliquer sur le fait et/ou sur les dimensions associées ;
- <**order by**> est également optionnelle et permet de classer les valeurs des paramètres.

Exemple. Supposons que des décideurs souhaitent analyser les montants moyens de location des véhicules de marque Peugeot en fonction des régions de location et des années. La requête OLAP-SQL permettant de répondre à ce besoin est la suivante :

```
select avg(montant)                                  => Mesure choisie
according to rows region of AGENCES with H_GEO,
             columns annee of TEMPS with H_AN        => Mode d'affichage
from LOCATIONS                                       => Fait étudié
where VEHICULES.marque='Peugeot' ;                   => Restriction
```

4.1.2 Complétude du langage assertionnel

Le tableau suivant permet de préciser pour les opérations algébriques du noyau minimum sa traduction en langage assertionnel.

Opérateur	Traduction OLAP-SQL
DrillDown	Ajouter un (ou plusieurs) paramètre(s) de plus bas niveau (suivant la hiérarchie courante) dans la clause <according to>
RollUp	Supprimer un (ou plusieurs) paramètre(s) dans la clause <according to>
Select	Spécification d'un prédicat dans la clause <where>
AddM /DelM	Modification de la clause <Select>
DRotate	Modification de la clause <according to>
Switch	Spécification de l'ordre dans la clause <order by>
Nest	Spécification de l'ordre dans la clause <according to>

Figure 41 : Traduction des opérateurs du noyau minimum en OLAP-SQL

Le tableau suivant permet de préciser pour chacune des opérations algébriques de second niveau sa traduction en langage assertionnel.

Opérateur	Traduction OLAP-SQL
HRotate	Modification de la clause <according to> au niveau du <with>
FRotate	Spécification d'un nouveau fait dans la clause <from> avec adéquation de l'ensemble des autres clauses
Plot	Préciser ces attributs dans la clause <according to>
Order	Spécification de l'ordre dans la clause <order by>
Unselect	Modification ou suppression d'un prédicat dans la clause <where>

Figure 42 : Traduction des opérateurs de second niveau en OLAP-SQL

4.2 LANGAGE DE DEFINITION DE OLAP-SQL

En plus de la commande d'interrogation SELECT, notre langage OLAP-SQL permet de définir, de contrôler et de manipuler une base de données en constellation. Dans cette section, nous nous focalisons sur la définition (LDD). Les principales caractéristiques de notre proposition sont :

- une définition aisée et adéquate des composants d'une base multidimensionnelle (fait, dimension, hiérarchie) en faisant abstraction des tables d'implantation,
- une définition facilitée des relations faits/dimensions masquant la définition des jointures implicites entre les dimensions et les faits,
- une définition explicite de multi-hiérarchies sur chaque dimension.

Pour la création d'un schéma nous avons proposé l'extension de la commande Create de SQL comme suit :

Commande	Commentaire
create dimension	Cette commande de création des dimensions s'inspire de la syntaxe introduite dans Oracle 8i. Cependant, nous offrons la possibilité de définir et de valuer simultanément les paramètres au travers d'extraction d'attributs de tables existantes (clause <as>)
create fact	Cette commande permet de définir les mesures ainsi que les dimensions en fonction desquelles les mesures sont analysées.

Figure 43 : commandes de création de OLAP-SQL

Pour modifier la structure des dimensions, des hiérarchies et des faits nous proposons une extension de la commande "Alter" en la déclinant en "Alter fact", "Alter dimenion" et "Alter hierachy". Il est possible d'effectuer les modifications suivantes :

- associer/dissocier une dimension à un fait ,
- ajouter, modifier ou supprimer un attribut (paramètre ou mesure) à une dimension, une hiérarchie ou un fait ,
- ajouter ou supprimer une hiérarchie à une dimension.

Le tableau suivant présente de manière succincte les combinaisons possibles

	Alter fact	Alter dimension	Alter hierarchy
connect	✓		
disconnect	✓		
add attribute	✓	✓	✓
drop attribute	✓	✓	✓
modify attribute	✓	✓	✓
add hierarchy		✓	
drop hierarchy		✓	

Figure 44 : déclinaison des commandes "Alter"

Les commandes de suppression des faits et des dimensions sont modélisées au travers de l'extension de "Drop".

4.3 LANGAGE DE CONTROLE DE OLAP-SQL

Les informations stockées dans une base multidimensionnelle sont manipulées par différents utilisateurs (administrateurs, décideurs de différents niveaux hiérarchiques). Nous souhaitons que OLAP-SQL contiennent également des commandes de contrôle permettant une définition précise et flexible adaptée à une politique de sécurité sur les données. OLAP-LCD n'a pas pour vocation de spécifier des droits sur les tables d'implantation mais s'oriente vers la définition de droits sur les types d'analyse (en fonction de sujets, d'axes d'analyse, de perspectives sur les analyses).

Notre contribution majeure réside en la proposition de commandes permettant d'affecter des droits sur les composants d'une constellation ou de révoquer ces droits. La spécificité d'OLAP-LCD est de définir de manière uniforme des droits sur chacun des composants d'un schéma en constellation (fait, dimension, hiérarchie, paramètre). Les droits sont associés aux commandes de sélection (SELECT) et de définition du schéma (ALTER, CREATE, DELETE, DROP, INSERT, UPDATE).

OLAP-LCD permet aux administrateurs d'autoriser différentes analyses sur une même base multidimensionnelle. Plus précisément, la structure du schéma en constellation sert de support à la définition de droits utilisateurs. Ainsi,

- l'accès à un fait autorise un sujet d'analyse,
- l'accès à une dimension autorise un axe d'analyse,
- l'accès à une hiérarchie donne une perspective d'analyse particulière,
- l'accès à une mesure et/ou un paramètre donne plus ou moins d'information.

La mise en place de droits sur un composant donne, par transitivité, les mêmes droits sur les éléments le composant. Autrement dit,

- un droit sur une hiérarchie permet d'obtenir le même droit sur les paramètres,
- un droit sur une dimension permet d'obtenir le même droit sur les hiérarchies de la dimension et les paramètres,

- un droit sur un fait permet d'obtenir le même droit sur les dimensions associées, les hiérarchies et les paramètres.

Ces droits s'appliquent à des utilisateurs ou à des groupes d'utilisateurs. Ainsi, il est possible de rendre visible certains composants à des utilisateurs spécifiques tandis que d'autres utilisateurs ne pourront disposer que de certains composants limitant ainsi leurs analyses. Les droits définis au niveau d'un groupe sont répercutés à l'ensemble des utilisateurs du groupe.

Ces droits sont définis au travers d'une extension de la commande "Grant" et "Revoke". La syntaxe précise de ces commandes est présentée dans [Ravat et al., 2002].

4.4 LANGAGE DE MANIPULATION DE OLAP-SQL

Le langage de manipulation de OLAP-SQL permet à l'administrateur d'insérer (insert), de supprimer (delete) et de modifier (update) les valeurs des faits ou des dimensions en faisant abstraction des tables d'implantation. Les commandes de modification et de suppression sont facilement appréhendables par l'administrateur car le fonctionnement est similaire aux commandes SQL définies pour les tables relationnelles. Cependant, la commande d'insertion que nous proposons est adaptée au contexte multidimensionnel :

- l'insertion de données dans une dimension est réalisée au travers d'une saisie explicite des valeurs ou par extraction de données source à l'aide d'une requête (cette requête permet de recopier des valeurs issues de bases de données externes dans la base multidimensionnelle) ;
- l'insertion de données dans un fait décharge l'administrateur de la gestion des références internes aux dimensions.

L'insertion de valeurs dans une dimension et un fait s'effectue respectivement avec les commandes "Insert into dimension" et "Insert into fact". La modification et/ou la suppression des dimensions et des faits s'effectuent avec un dérivé de la commande "Update".

La syntaxe précise de OLAP SQL est donnée dans [Ravat et al., 2002].

5 BILAN ET PERSPECTIVES

Dans le troisième chapitre, nous avions explicité les différents concepts du modèle multidimensionnel. Notamment, ce modèle comportait le concept de Table Multidimensionnelle (TM) permettant de visualiser les données d'une constellation lors des analyses décisionnelles. Ce chapitre a permis de présenter les travaux que nous avons menés dans le domaine des langages de manipulation de données au travers des TM. Notre objectif était d'apporter une solution complète en proposant un langage algébrique, un langage graphique et un langage assertionnel.

5.1 BILAN SUR LES LANGAGES DE MANIPULATION

Dans le domaine des algèbres OLAP, il n'existe de proposition stable et reconnue par la communauté scientifique. De plus, aucun des travaux n'a proposé une étude exhaustive des opérations de manipulation multidimensionnelle.

Nos travaux visent à répondre à cette problématique. Notamment, notre souhait était d'apporter une solution pour une algèbre orientée décideur proposant à la fois des structures conceptuelles adaptées (TM) et des commandes algébriques facilement interprétables par ces derniers. Dans un premier temps, nous avons défini un opérateur de construction permettant de créer une TM à partir d'un schéma multidimensionnel. Dans un second temps, à l'instar du modèle relationnel, nous avons proposé un **noyau minimum** permettant d'effectuer les manipulations décisionnelles essentielles. Ces manipulations permettent de paramétrer une

analyse (opérations de forage, rotation et sélection), de spécifier les options de présentation (opérations de permutations de paramètres et d'agrégation) et de transformer une TM (opérations de modification de fait ou de dimension). Nous avons complété ce noyau minimum par l'ajout de **commandes de second niveau** permettant de simplifier l'écriture d'une requête et d'optimiser les traitements réalisés par un SGBD lors de leurs utilisations. Pour terminer, nous avons proposé un ensemble d'**opérateurs ensemblistes flexibles** permettant d'effectuer des unions, des différences et des intersections entre TM pour en générer une nouvelle. Ces opérations ensemblistes permettent de faciliter les corrélations inter-tables multidimensionnelles inhérentes à un processus d'analyse. De part sa propriété de **fermeture**, cette algèbre permet de répondre à des requêtes complexes en combinant des opérations algébriques. Afin de représenter la puissance d'expressivité de nos travaux, nous présentons le **tableau comparatif** suivant :

Opérations \ Travaux de Recherche		(Grouping Algebra) Li & Wang 1996	Agrawal *et al.*, 1997	Gyssens & Lakshmanan, 1997	(MD) Cabibbo & Torlone 1997, 1998	Lehner, 1998	Datta & Thomas, 1999	Pedersen *et al.*, 2001	(YAM²) Abelló *et al.*, 2003	(GMD) Franconi *et al.*, 2004	Notre proposition
Forage	Niveau plus détaillé	Roll ,Cube	Join			DrillDown[2], Split	(Push, Pull) + Join		DrillDown		DrillDown
	Niveau moins détaillé	Roll, Aggregation	Merge	Summerization	RollUp, Aggregation	RollUp, Merge, Aggregation	-	Aggregation	RollUp		RollUp
Sélection	Valeurs factuelles			Slice (Selection)	Selection		Restriction + Pull		Dice, Projection		Select
	Valeurs dimensionnelles		Restriction	Dice (Selection)	Selection		Restriction	Selection		Slice, Multi-Slice[3]	Select
Rotation	Fait								DrillAcross		FRotate
	Dimension								ChangeBase		Rotate
	Hiérarchie										HRotate
Modification de fait	Ajout d'une mesure		Projection	Projection						Derived measures	AddM
	Suppression d'une mesure		Projection	Projection							DelM
Modification de dimension	Réduction de dimension	Cube Aggregation	Projection, Destroy-Dimension		Simple Projection		Partition	Projection		Projection	Display
	Push		Push	Fold[4]			Push				Push
	Pull		Pull	Unfold			Pull				Pull
Ordonnancement	Order			Classification							Switch, Order
	Imbrication	Transfer									Nest
Opérateurs binaires	Union	Union[6]	Union[6]	Union[6]			Union[5]	Union[5]	Union[6]	Union[6]	Union
	Intersection		Intersection[6]	Intersection[6]			2 Difference			Intersection[6]	Intersect
	Différence		Difference[6]	Intersection[6]			Difference[5]	Difference[5]		Intersection[6]	Difference
	Jointure	RC-Join (Relation vers dimension)	join cubes	join cubes	Join[1]		Cartesian product + Restriction	Identity-based Join, Group		join cubes	Union +Sélection
Structure du modèle		Cube	Cube	2D-Table (TM)	MD (f-table)	MD	Cube	MD	Cube	Cube	TM
Autres opérations		Add dimension		cartesian product	cartesian product		Cartesian product				Plot
Autres langages				langage de calcul	langage graphique et "query calculus"				traduction en SQL		langage graphique et langage assertionnel

MD=Multidimensionnel; [1]=sans restriction; [2]=pas de conservation de hiérarchies; [3]=spécifié sur une zone; [4]=push généralisé; [5]=sur dimensions communes; [6]=cubes identiques seulement;

Figure 45 : Comparatif des opérateurs algébriques

Une telle algèbre permet d'expliciter les opérations effectuées par un décideur et de les traduire facilement en algorithmes d'exécution par un SGBD. Or, les décideurs souhaitent des langages moins abstraits pour manipuler une BDM. Pour répondre à leurs besoins, nous avons proposé un **langage graphique**. Contrairement aux propositions du marché, l'avantage majeur de notre solution est de p**résenter une BDM au travers d'un graphe représentant le schéma multidimensionnel**. Ce graphe sert de support à l'élaboration incrémentale de requêtes décisionnelles. Cette solution lui permet de faire complètement abstraction des opérations associées à ses manipulations graphiques. L'autre avantage de notre proposition est que ce **langage graphique est complet** au regard de l'algèbre que nous avons proposée précédemment. Ce langage est implanté dans le cadre d'un prototype R-OLAP (cf. chapitre 6) et toute manipulation graphique est traduite en commandes SQL.

Dans le cadre des langages, notre dernière proposition visait à offrir aux utilisateurs avertis un **langage assertionnel** pour créer une TM. Le langage assertionnel proposé permet non seulement d'interroger les données multidimensionnelles mais également de définir et contrôler ces données. Notre langage est basé sur une **extension de SQL**. De plus, au travers de la commande SELECT étendue, ce langage présente deux avantages : définition uniforme des TM

et **complétude** au regard de l'algèbre définie précédemment. Ce langage a également fait l'objet d'une implantation (cf. chapitre 6) ; toute commande SELECT sur une TM est décomposée en plusieurs commande SELECT sur les tables de la BDM.

Tous ces langages ont été définis sur le noyau de base de notre modèle multidimensionnel. Pour le modèle intégrant les contraintes, une extension de l'algèbre a déjà été proposée [Ghozzi, et al., 2004 ; Ravat et al., 2005a]. Par contre, pour la manipulation d'un schéma intégrant des versions, nous sommes en cours de finalisation dans la proposition d'une algèbre.

5.2 PRODUCTION SCIENTIFIQUE

Ces travaux sont le résultat du co-encadrement de deux thèses. En complément de leurs travaux sur la modélisation multidimensionnelle, les thésards Faiza Ghozzi et Ronan Tournier ont respectivement apporté des solutions pour manipuler des données contraintes sémantiquement [Ghozzi, 2004] et des données documentaires [Tournier, 2007]. Notamment, ces derniers travaux ont apporté des solutions pour l'agrégation de données textuelles.

Ces travaux ont également donné lieu à l'encadrement de 6 masters recherche. A. Tahi et L. Bouzguenda ont validé le principe de l'interrogation multidimensionnelle incrémentale [Tahi, 2005] au travers d'une table multidimensionnelle [Bouzguenda, 2002]. Ces travaux ont été complétés par l'implantation du module d'interrogation assertionnelle [Annoni, 2003], du module d'interrogation graphique [Tournier, 2003], du module d'interrogation assertionnelle à base de contraintes sémantiques [Le Thi, 2003] et de la fusion de deux TM [Rouhaud, 2005]. Ces travaux sont également le fruit de l'encadrement d'un magistère SIC (Système d'Information et de Communication) de l'INI (Institut National de Formation en Informatique d'Alger) sur les langages graphiques pour bases de données temporelles (S. Hafyane).

Ces travaux sont un des points fondateurs de notre participation au projet IAPA[14] qui est en cours de finalisation. Ce projet fait intervenir plusieurs équipes de l'IRIT et vise à proposer une infrastructure d'accès, de partage et d'analyse de données biomédicales. Dans ce cadre, nous avons la responsabilité du lot relatif à la modélisation multidimensionnelle des données ainsi qu'à leurs analyses.

D'un point de vue publications, nous pouvons citer les références suivantes[15] :

- 1 article dans une revue internationale : IJDWM [Ravat et al., 2007a], revue de référence dans le domaine du décisionnel,
- 2 articles dans des conférences internationales : ICEIS'07 [Ravat et al., 2007b], ADBIS'07 [Ravat et al., 2007d],
- 1 article dans une revue nationale : RSTI-ISI [Ravat et al., 2002],
- 2 articles dans des conférences nationales : INFORSID'06 [Ravat et al., 2006b], EGC'2005 [Ravat et al., 2005a].

5.3 PERSPECTIVES

Nos travaux présentent l'avantage d'intégrer trois composants complémentaires : une algèbre (avec un noyau minimum, des opérateurs de second niveau et des opérateurs binaires), un langage graphique et un langage assertionnel.

[14] Le contenu de ce projet sera explicité dans le chapitre 6.
[15] Le contenu de chaque article est résumé dans le chapitre 6.

Au niveau de l'algèbre, nous avons proposé l'extension des opérations "emblématiques" de rotation et de forage pour les schémas multidimensionnels contraints sémantiquement [Ghozzi, et al., 2004] et personnalisés [Ravat et al., 2007f]. Une première perspective consiste à compléter ce travail en étudiant de manière exhaustive l'ensemble des opérateurs proposés dans notre algèbre. A plus long terme, il faudrait proposer une algèbre complète pour les schémas multidimensionnels à base de versions. Cette extension constitue un travail de longue haleine notamment pour les manipulations décisionnelles faisant intervenir différentes versions de dimension ou d'étoile.

De plus, les extensions proposées se sont uniquement concentrées sur les opérations algébriques. Une autre perspective de travail consiste à proposer une extension des langages graphique et assertionnel précédemment définis.

CHAPITRE V : DEMARCHE DE CONCEPTION

PLAN DU CHAPITRE

1 INTRODUCTION A LA DEMARCHE DE CONCEPTION DE SYSTEME DECISIONNEL 107

2 CONCEPTION DE SCHEMAS MULTIDIMENSIONNELS CONTRAINTS 107
2.1 Problématique 108
2.2 Etapes de la démarche mixte 108
2.3 Approche descendante 109
2.3.1 Collecte des données 109
2.3.1.1 Requêtes-types et questionnaires 109
2.3.1.2 Règles de gestion 110
2.3.2 Spécification des besoins 110
2.3.3 Formalisation des besoins 111
2.4 Approche ascendante 111
2.4.1 Détermination des faits 112
2.4.2 Détermination des dimensions 113
2.4.3 Définition de la granularité de l'analyse 113
2.4.4 Hiérarchisation des dimensions 113
2.4.5 Expression des contraintes 114
2.5 Confrontation et bilan 114

3 CONCEPTION D'UN SYSTEME D'AIDE A LA DECISION 115
3.1 Problématique 115
3.2 Démarche du Trident Décisionnel 116
3.3 Analyser le SAD 119
3.3.1 Caractériser les groupes d'acteurs 119
3.3.2 Analyser les besoins tactiques 119
3.3.3 Analyser les besoins stratégiques 121
3.3.4 Analyser les besoins systèmes 121
3.3.5 Les confrontations 121
3.4 Concevoir le SAD 121
3.4.1 Architecture modulaire 122
3.4.2 Concevoir les modules 123
3.5 Principes de réutilisation 124
3.6 Implantation au sein de la société I-D6 126

4 BILAN ET PERSPECTIVES 128

4.1 CONCEPTION D'UN MAGASIN DE DONNEES MULTIDIMENSIONNELLES 128
4.2 Conception d'un système d'aide à la décision 129
4.3 Production scientifique 129
4.4 Perspectives 130

1 INTRODUCTION A LA DEMARCHE DE CONCEPTION DE SYSTEME DECISIONNEL

L'objectif de nos travaux est d'apporter des solutions méthodologiques pour la conception et le développement de Systèmes d'Aide à la Décision (SAD). Dans les chapitres précédents, nous avons présenté l'architecture des SAD basée sur la dichotomie entrepôt et magasins de données, des modèles pour les deux espaces de stockage et des langages de manipulation de données multidimensionnelles. Nous devons compléter ces propositions par une démarche de conception.

A l'heure actuelle, il n'existe pas de méthode de développement de SAD reconnue [Rizzi et al., 2006]. La définition d'une telle méthode nécessite encore de nombreux travaux. Cette situation est liée au fait que le développement de SAD repose sur les spécificités suivante :

- les données de ces SAD reposent sur une intégration de sources hétérogènes internes et/ou externes à l'organisation ;
- l'expression des besoins est effectuée par des décideurs de différents niveaux hiérarchiques (stratégiques, tactiques), de différentes divisions de l'organisation représentant des intérêts de domaines spécifiques voire antagonistes entre eux ;
- les données sont stockées dans différents espaces de stockage ayant des vocations différentes (préparation des données pour un ED, présentation des données pour un MD) et reposant sur des modèles spécifiques. Il faut également définir les processus d'extraction, de transformation et de chargement des données entre ces différents espaces de stockage.

De part ces caractéristiques, il était impossible d'adapter une méthode de conception d'application transactionnelle au monde du décisionnel. Notre problématique est donc de proposer en complément de nos modèles, une démarche de conception adaptée. Après une étude des différents travaux, nous pouvons remarquer que ces derniers se sont centrés essentiellement sur les processus de développement d'un seul composant, à savoir les magasins de données multidimensionnelles [Cabibbo & Torlone, 1998 ; Cabibbo & Torlone, 2000 ; Golfarelli & Rizzi, 1998 ; Moody & Kortink 2000 ; Tsois et al., 2001 ; Prat & Akoka, 2002 ; Prat et al., 2006 ; Carneiro & Braymer, 2002 ; Luján-Mora & Trujillo, 2003 ; Luján-Mora & Trujillo, 2003].

Pour répondre à cette problématique, nous avons procédé en deux temps. Tout d'abord, nous nous sommes concentrés sur la particularité des SAD, à savoir, la conception de magasins multidimensionnels. Plus précisément, notre souhait est de proposer une démarche pour le développement de Base de Données Multidimensionnelles (BDM) intégrant des contraintes sémantiques afin d'assurer la validité des données et d'interdire des analyses décisionnelles comportant des incohérences. Dans un second temps, nous avons étendu nos travaux afin de proposer une démarche pour la conception de systèmes décisionnels complets. Afin de valider nos propositions, nous les avons confronté aux besoins d'une société de service spécialisée dans le décisionnel. Chacun de ces points est étudié dans les sections suivantes.

2 CONCEPTION DE SCHEMAS MULTIDIMENSIONNELS CONTRAINTS

L'objet de cette section est la proposition d'une méthode de conception de magasin de données multidimensionnelles intégrant un ensemble de contraintes sémantiques. La conception d'une base de données décisionnelles doit répondre de façon spécifique aux besoins de l'entreprise et apporter les réponses adéquates aux requêtes des décideurs. Aussi, la définition d'une méthode de conception d'une telle base est cruciale [Bonifati et al, 2001 ; Kimball & Ross, 2002].

2.1 PROBLEMATIQUE

La méthode de conception que nous souhaitons proposer repose sur les formalismes et les concepts énoncés dans le troisième chapitre. Notamment, nous souhaitons proposer une méthode pour la conception de BDM contenant des contraintes sémantiques permettant de vérifier la validité des données et garantissant la cohérence des analyses décisionnelles.

Dans la littérature, les méthodes sont classées en trois catégories : ascendantes, descendantes et mixtes. Comme indiqué en section 3.4 du premier chapitre, les **méthodes mixtes** combinent les deux méthodes précédentes et essayent de combler les lacunes de chacune d'elles. Pour ces raisons, nous souhaitons **positionner nos travaux dans cette catégorie**. Cette solution présente l'avantage d'intégrer les besoins initiaux des décideurs éventuellement amendés en fonction des données contenues dans l'entrepôt de données sources. Par opposition aux travaux actuels, notre proposition doit permettre de définir précisément les étapes liées à la prise en compte des besoins des décideurs et à leur transformation en schéma multidimensionnel. De plus, contrairement aux autres propositions relatives à la démarche mixte, nous ne souhaitons pas seulement analyser les données sources pour construire un schéma final mais bien détecter les différents sujets et axes d'analyse contenus dans les sources pour construire un schéma multidimensionnel. Ainsi dans notre proposition, l'approche orientée utilisateurs (descendante) permet d'identifier un schéma des données "idéal" alors que l'approche orientée données (ascendante) fournit un schéma "candidat". La phase de confrontation permet définir le schéma final à partir de ces deux propositions. Dans les sections suivantes, nous présentons les différentes étapes de cette démarche mixte.

2.2 ETAPES DE LA DEMARCHE MIXTE

La méthode proposée permet de concevoir une BDM selon les trois niveaux d'abstraction recommandés par l'ANSI/X3/SPARC (conceptuel, logique et physique). Au niveau conceptuel, nous obtenons un schéma multidimensionnel contraint qui sera traduit en schéma R-OLAP au niveau logique, puis en un schéma spécifique à un SGBD au niveau physique.

Dans cette section, nous présentons uniquement la phase conceptuelle de notre méthode de conception. Les phases logique et physique sont présentées dans le cadre de notre système d'aide à la conception de systèmes décisionnels (chapitre VI).

La mise au point d'une démarche mixte nécessite la définition en amont du cadre général de l'application décisionnelle. En effet, nous ne souhaitons pas être confrontés au problème de divergence de schéma des résultats des approches descendantes et ascendantes dans un contexte de SAD multi-domaines. De plus, la détermination de tous les sujets et axes d'analyse à partir des sources s'avère fastidieuse et irréaliste. Pour résoudre ce problème, nous avons proposé une étape de définition du ***domaine de l'analyse*** en amont des démarches descendante et ascendante.

Dans le schéma suivant, vous retrouvez les différentes étapes de notre démarche qui seront explicitées dans les sections suivantes :

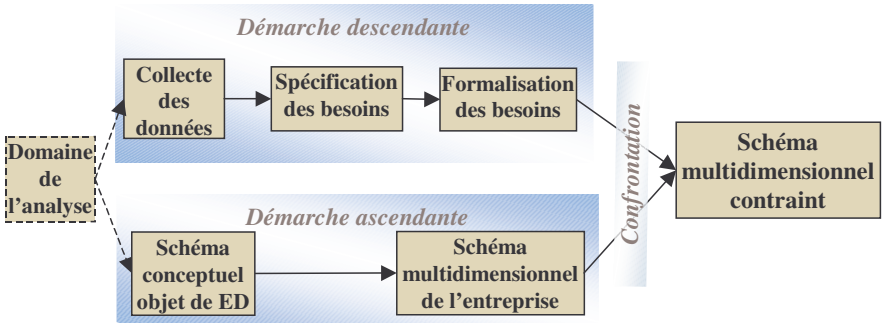


Figure 46 : Etapes de la méthode de conception

2.3 APPROCHE DESCENDANTE

L'objectif de cette démarche est de concevoir un schéma dimensionnel en se basant sur les besoins des décideurs et sur les règles de gestion relatives aux données décisionnelles. Cette démarche se base sur trois étapes : (1) collecte des données, (2) spécification des besoins et (3) formalisation des besoins. Les sections suivantes explicitent chacune de ces étapes.

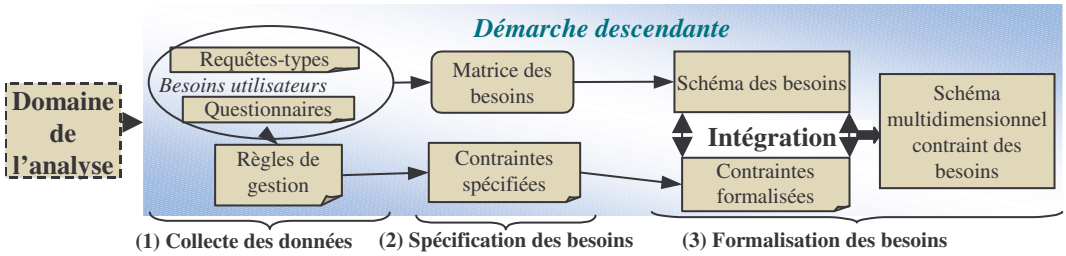


Figure 47 : Démarche descendante

2.3.1 Collecte des données

Dans cette étape, nous déterminons les besoins décisionnels initiaux en définissant les types d'informations susceptibles d'intéresser chaque groupe de décideurs. Nous procédons alors à :

- la collecte des **requêtes-types** pertinentes en interviewant les décideurs,
- la mise au point d'un **questionnaire** permettant de mieux caractériser les besoins,
- l'analyse des données décisionnelles, des interviews et des questionnaires afin de dégager les **règles de gestion** appliquées au domaine.

2.3.1.1 Requêtes-types et questionnaires

Les requêtes-types sont élaborées à partir des anciens rapports d'analyse ou en interviewant les décideurs. Ces requêtes-types sont exprimées à l'aide d'un pseudo-langage facilitant la définition des besoins et permettant de guider le concepteur dans la définition ultérieure d'un schéma multidimensionnel. Ce pseudo-langage est basé sur les trois clauses suivantes :

- la clause ***Analyser*** répond à la question ***Quoi ?*** en spécifiant les données que les décideurs souhaitent analyser ;
- la clause ***En fonction*** répond aux questions ***Qui ? Où ?*** et ***Quand ?*** en précisant les paramètres de l'analyse des données décrites par la clause ***Analyser*** ;
- la clause ***Pour*** répond aux questions ***Pour Qui*** ou ***Quelles données*** ? en précisant les restrictions sur les données à analyser.

Exemple. Nous souhaitons spécifier un magasin multidimensionnel permettant d'analyser les locations de véhicules. L'analyse des rapports d'activité existants et une première interview des décideurs ont permis de spécifier six requêtes-types (***R1*** à ***R6***). La figure suivante présente un exemple de rapports d'activité que nous avons collecté pendant l'analyse. A partir de ces rapports, nous avons extrait les requêtes-types ***R2*** et ***R3***.

Listing des montants des locations par Mois et par Agence :

Mois Location	Agence	Montant Location
Janvier-00	Ag_tlse	90
Janvier-00	Ag_bord	120
Juin-00	Ag_tlse	70
juin-00	Ag_Dallas	110
Juin-00	Ag_Gren	200
Janvier-01	Ag_tlse	250
…	..	…

Listing des montants des locations par Ville :

Ville	Montant Location
Toulouse	410
Bordeaux	120
Grenoble	200

R1 : ***Analyser*** le montant des locations
En fonction des mois et des véhicules
Pour les véhicules de type sport.

R2 : ***Analyser*** le montant des locations
En fonction des mois et des agences

R3 : ***Analyser*** le montant des locations
En fonction des villes

R4 : ***Analyser*** le chiffre d'affaire des employés
En fonction des mois
Pour l'employé Paul et l'année 2002.

R5 : ***Analyser*** les marques des véhicules
En fonction de la durée de location
Pour l'état de Floride.

R6 : Analyser les montants des locations
En fonction des véhicules
Pour la marque Peugeot.

Figure 48 : Un exemple de rapport d'activité

Le questionnaire est un outil de collecte complémentaire de données dans lequel le décideur est guidé au travers d'un ensemble de questions dans le but d'obtenir des informations précises, nécessaires à la modélisation multidimensionnelle. Les questionnaires permettent de modifier et éventuellement, d'ajouter de nouveaux éléments aux requêtes-types (indicateurs, paramètres, hiérarchie…).

2.3.1.2 Règles de gestion

Les règles de gestion régissent le système d'information et permettent d'assurer le bon fonctionnement des systèmes opérationnels sources. Pour l'élaboration d'un SAD, ces règles seront extraites des systèmes opérationnels. Elles peuvent être spécifiées dans la documentation, les anciens rapports d'activité, les questionnaires ou les interviews des décideurs. Ces règles permettent notamment de spécifier la hiérarchisation des données.

Exemple. Les agences de location de véhicules sont localisées soit en France, soit aux Etats-Unis. La description géographique de ces agences repose sur une description différente pour la France (ville, département, pays) et pour les Etats-Unis (ville, état).

2.3.2 Spécification des besoins

L'étape de spécification permet d'analyser et d'organiser les données des requêtes types et les règles de gestion définies précédemment.

Premièrement, cette étape permet de traduire les requêtes types en une matrice des besoins spécifiant les relations entre les indicateurs et les paramètres d'analyse. Cette matrice est une matrice carrée dont les lignes et les colonnes regroupent toutes les propriétés. Chaque case cochée (X) indique que la propriété en ligne est analysée en fonction de celle en colonne. Après simplification de la matrice (suppression des lignes et des colonnes vides), la matrice contient en lignes les mesures et en colonnes les paramètres d'analyse.

Exemple. L'étude des 6 requêtes-types présentées dans l'exemple précédent permet de construire la matrice des besoins simplifiée suivante :

Matrice des besoins

Paramètres / Indicateurs	Ville	Région	Etat	Année	Mois	Id_Emp	Id_Client	Immat	Type	Marque
Mt locations	x	x	x	x	x		x	x	x	X
Nb jours	x	x	x	x	x		x	x	x	x
CA				x	x	x				

Figure 49 : Exemple de matrice des besoins simplifiée

Dans un second temps, cette étape de spécification permet de traduire les règles de gestion définies dans l'étape de collecte de besoins en des contraintes facilement intégrables dans un schéma multidimensionnel. Une règle de gestion peut donner lieu à plusieurs contraintes appliquées à différentes données multidimensionnelles. La spécification de ces contraintes peut s'effectuer en langage naturel ou à l'aide d'une extension du langage des contraintes objet OCL[16] que nous appelons LCD (Langage de Contraintes multiDimensionnelles [Ghozzi et al., 2005]).

2.3.3 Formalisation des besoins

Après avoir collecté et spécifié les besoins des décideurs, nous réalisons dans cette étape la formalisation de ces besoins sous forme d'un schéma multidimensionnel contraint.

Dans un premier temps, nous spécifions les structures d'un schéma multidimensionnel en extrayant l'information contenue dans la matrice des besoins. Outre le fait que cette représentation multidimensionnelle soit adaptée pour la définition des besoins décisionnels [Kimball et al., 2005], elle permettra d'offrir un cadre facilitant l'étape de confrontation entre les résultats des approches descendantes et ascendantes. La construction d'un schéma multidimensionnel s'effectue à partir d'une matrice des besoins comme suit :

- ***définition des faits***. La définition des faits peut être réalisée de manière automatique en regroupant les mesures analysées au travers de paramètres identiques ;
- **définition des dimensions.** Cette étape réalise le regroupement automatique des paramètres caractérisant les mêmes mesures en une dimension, l'éventuel enrichissement du contenu de ces dimensions par l'ajout d'attributs faibles, et enfin, la spécification du niveau d'analyse le plus fin pour chaque dimension (granularité d'analyse) ;
- ***définition des hiérarchies***. Cette étape consiste à organiser les paramètres de chaque dimension dans des hiérarchies. Cette hiérarchisation se base sur la connaissance du domaine ;
- ***définition du schéma multidimensionnel***. Cette étape permet une affectation automatique des dimensions aux faits à partir de la matrice des besoins.

Dans un second temps, nous intégrons les contraintes intra et inter-dimensions entre les hiérarchies de dimensions. Ces contraintes sont directement traduites de la liste définies lors de l'étape précédente.

2.4 APPROCHE ASCENDANTE

Afin de tenir compte des informations contenues dans la source de données (l'entrepôt pour notre cas), nous proposons une approche ascendante incrémentale. Cette démarche part du schéma conceptuel de l'entrepôt de données historisées pour construire le schéma multidimensionnel contraint d'un magasin de données [Bret & Teste, 1999]. Le concepteur doit détecter les différents centres d'intérêt de l'organisation en analysant le schéma de l'entrepôt. Nous considérons que l'approche ascendante est réalisée en parallèle de la démarche descendante. Toutefois, le concepteur peut, au niveau de la phase ascendante, tenir compte d'informations collectées lors de la démarche descendante.

L'objectif de cette démarche est de concevoir le s**chéma multidimensionnel de l'entreprise** en se basant sur le schéma conceptuel de l'entrepôt de données et sur le domaine de l'analyse. Nous avons subdivisé notre démarche descendante en 6 étapes.

[16] http://www.omg.org/docs/formal/03-03-13.pdf

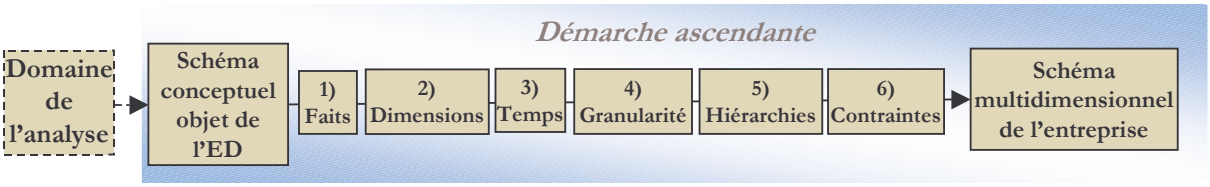


Figure 50 : Etapes de la démarche ascendante

2.4.1 Détermination des faits

La détermination des sujets de l'analyse permet de dresser la liste des faits du modèle dimensionnel. Selon notre démarche, un fait est projeté à partir d'une Classe Représentative de l'entrepôt. Une Classe Représentative (CR) décrit un événement qu'un décideur veut analyser (vente, achat, …) et contient des valeurs numériques servant de support à l'élaboration des mesures d'activité. De plus, une CR est une classe fréquemment mise à jour dans l'entrepôt par opposition aux autres classes plutôt statiques.

Définition. Une Classe Représentative (CR)

- est fréquemment mise à jour dans un entrepôt,
- décrit un événement qu'un décideur veut analyser, et
- contient des attributs numériques servant de support à l'élaboration des mesures.

Tout fait est construit à partir d'une CR. Les mesures du fait sont donc obtenues en appliquant une fonction de calcul sur un ou plusieurs attributs de la classe représentative.

Exemple. Supposons que l'on possède l'entrepôt de données représenté ci-après. Nous pouvons repérer deux environnements et différentes classes entrepôts. Nous souhaitons analyser l'activité des agences de locations. L'étude du schéma permet de détecter une classe représentative. Cette CR est la classe "location" et elle possède deux attributs susceptibles de devenir des mesures d'activités ("montant_loc" et "duree")

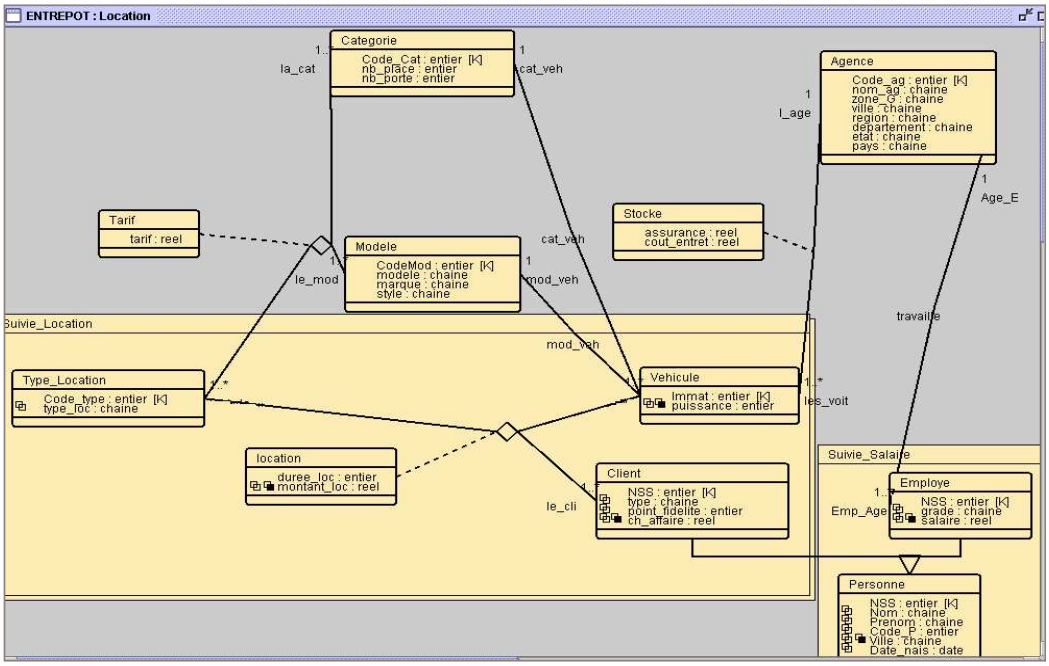


Figure 51 : Exemple d'entrepôt source

2.4.2 Détermination des dimensions

Les dimensions sont élaborées à partir des classes entrepôts reliées à la classe représentative, appelées **classes déterminantes** [Bret & Teste, 1999]. La détermination de ces classes est réalisée automatiquement en suivant le principe de dépendance fonctionnelle entre classes.

Définition. Une classe Ci est déterminante d'une classe CR noté Ci => CR si

- Ci = CR,
- Ci hérite de CR,
- Ci représente une classe d'association et CR entre dans la formation de l'association,
- Ci est reliée à CR par une relation d'association monovaluée (X, 1) ou
- Ci est reliée à CR par une relation de composition de type (1, x) si Ci est composée de CR ou (x, 1) si Ci compose CR.

La contrainte sur les cardinalités des relations d'association et de composition permet de garantir l'unicité entre les valeurs d'une classe C_1 et les valeurs liées d'une classe C_2. Cette propriété est essentielle, car elle permet de relier dans le magasin les mesures issues de la classe représentative aux paramètres issus des classes déterminantes. On peut déterminer l'ensemble $Determin\ (CR) = \{CD_1, CD_2, \ldots, CD_m\}$ des classes déterminantes d'une classe représentative CR afin de spécifier l'ensemble des classes sources à partir desquelles peuvent être créées les dimensions.

Le principe de dépendance respecte la propriété de transitivité. Cette propriété permet de définir une classe C1 déterminante d'une classe C3 si la classe C1 est déterminante d'une classe C2 elle-même déterminante : si C1=>C2 et C2=>C3 alors C1=>C3.

Après constitution de cet ensemble, le concepteur sélectionne les classes déterminantes pour les transformer en dimensions susceptibles d'intéresser les décideurs du domaine d'analyse. Le concepteur peut toutes les choisir, en sélectionner certaines voire regrouper celles obtenues par transitivité en une seule. Les attributs des dimensions sont obtenus par application de fonctions sur un ou plusieurs attributs d'une classe déterminante. Notamment, les attributs de la dimension temporelle sont déterminés par application d'une fonction sur un même attribut origine.

2.4.3 Définition de la granularité de l'analyse

La quatrième étape consiste à déterminer le niveau de granularité le plus fin suivant lequel les données décisionnelles sont analysées. Cette étape détermine le paramètre de plus bas niveau pour chaque dimension. Chacun de ces paramètres représente le début d'une structure hiérarchique que nous allons définir dans l'étape suivante.

2.4.4 Hiérarchisation des dimensions

Les paramètres des dimensions sont organisés en une structure hiérarchique qui permet d'analyser les mesures à différents niveaux de granularité. La définition d'une hiérarchie est réalisée par la détection des **dépendances hiérarchiques** entre les paramètres d'une même dimension.

Définition. Une dépendance hiérarchique entre deux paramètres p_i et p_j implique que :

- toute valeur de p_i détermine de manière unique une valeur de p_j (dépendance fonctionnelle)
- p_j ne dépend pas fonctionnellement de p_i.

En se basant sur ce principe, nous pouvons définir différentes hiérarchies dans la même dimension. Au niveau de cette démarche, nous proposons de définir les hiérarchies les plus complètes possible en fonction du schéma de la source.

2.4.5 Expression des contraintes

Les contraintes sont des expressions qui précisent le rôle ou la portée d'un élément de modélisation (elles permettent d'étendre ou de préciser sa sémantique) [Doucet et al, 1996]. Nous avons proposé un ensemble de contraintes sémantiques pour représenter de manière précise la réalité et pour interdire les analyses non significatives. Nous avons proposé deux types de contraintes : intra et inter-dimensions. Elles permettent de définir des relations d'**exclusion**, d'**inclusion**, de **partition**, de **simultanéité** ou de **totalité** entre respectivement des hiérarchies d'une même dimension ou de deux dimensions.

La détection des contraintes intra-dimension est basée sur l'analyse des données de l'entrepôt. Par exemple, l'analyse des valeurs de la classe *Agence de location*, nous a permis de constater que, pour toutes les agences françaises, l'attribut *Etat* est nul et que, pour toutes les agences américaines, les attributs *Département* et *Région* sont nuls. Ce fait est exprimé au niveau de notre schéma conceptuel à l'aide de *la condition* d'*appartenance* aux hiérarchies. Dans un second temps, nous constatons que toutes les agences de la dimension font partie des agences françaises ou bien des agences américaines.

De la même manière que pour les contraintes intra, les contraintes inter-dimensions sont définies suite à l'analyse des instances des classes représentatives et des classes déterminantes de l'entrepôt. Par exemple, l'analyse des données de la classe représentative *Location* nous permet de constater que les instances de cette classe, reliées aux instances des agences françaises, ne sont pas associées aux instances de la classe véhicule organisées suivant une classification américaine.

2.5 CONFRONTATION ET BILAN

Après avoir conçu le schéma multidimensionnel des besoins en suivant la démarche descendante et le schéma multidimensionnel de l'entreprise en se basant sur l'entrepôt de données, nous procédons à l'intégration de ces deux schémas. Cette fusion est réalisée en confrontant les deux schémas afin d'enlever, d'ajouter ou de purifier quelques informations. L'intervention des décideurs au niveau de l'intégration est primordiale afin de tester la complétude du schéma multidimensionnel. Nous notons que l'intégration de schémas relève de plusieurs problématiques notamment au niveau de la sémantique (conflits de nom, de contexte, d'unité de mesure et de structure) [Kedad & Métais, 1999]. Ces problématiques ne sont pas traitées dans notre cas car nous considérons qu'au niveau de la définition du domaine de l'analyse, nous définissons un même dictionnaire de données et un même contexte.

Ainsi, l'intégration des données des deux schémas permet :

- ***une définition correcte de la granularité de l'analyse.*** Le choix d'un niveau de détail est très important dans la conception d'une base multidimensionnelle. Le choix d'une granularité trop fine risque d'augmenter la taille de la base multidimensionnelle en dérivant à partir de l'entrepôt des données détaillées non pertinentes pour l'analyse. Par contre, le choix d'une granularité moins fine ne permet pas d'analyser les données les plus détaillées. A ce niveau, le concepteur doit faire appel aux décideurs pour confronter leurs besoins exprimés au niveau de la démarche descendante aux granularités d'analyse détaillées obtenues lors de la démarche ascendante. Cette confrontation permet de découvrir des besoins non déclarés par les décideurs afin de déterminer une granularité adaptée aux besoins décisionnels ;

- ***l'épuration des données***. Les données extraites à partir des requêtes utilisateurs peuvent ne pas exister dans notre entrepôt de données. Dans ce cas, il est possible soit d'enlever ces données soit de les garder avec des valeurs vides en attendant d'avoir les informations nécessaires à leur instanciation ;
- ***l'ajout des données sources.*** C'est le concepteur qui analyse les données de l'entrepôt et détecte les informations susceptibles d'intéresser le décideur. Souvent, ce dernier ne connaît pas le contenu de l'entrepôt ni les informations susceptibles d'être analysées. Nous pouvons, par exemple, ajouter des paramètres d'analyse en détectant de nouvelles propriétés dans les dimensions définies dans notre exemple de schéma dimensionnel ;
- ***l'ajout des mesures calculées demandées par le décideur.*** Ces mesures nécessitent l'utilisation d'une règle de gestion qui permet de calculer la mesure à partir d'un ensemble de données entrepôt ;
- ***l'intégration de toutes les contraintes sémantiques.*** Ces contraintes proviennent soit de la démarche descendante (proposition du concepteur), soit de la démarche ascendante (analyse de données de l'entrepôt).

3 CONCEPTION D'UN SYSTEME D'AIDE A LA DECISION

Comme vu dans le premier chapitre, il existe différents modèles et différentes méthodes de développement de Système d'Aide à la Décision (SAD). Cependant, il n'existe pas de méthode de développement de SAD reconnue par les communautés scientifique et industrielle [Rizzi et al., 2006]. De plus, pratiquement tous ces travaux se centrent sur la modélisation multidimensionnelle des données. Cette section expose les solutions que nous avons apportées pour la conception de SAD complet. Cette proposition repose sur une extension des travaux que nous avons présentés dans la section précédente.

Après avoir exposé la problématique de recherche, nous étudierons la démarche proposée et nous expliciterons plus en détail les phases d'analyse et de conception.

3.1 PROBLEMATIQUE

80% des projets décisionnels n'atteignent pas les besoins des utilisateurs et 40% ne facilitent pas la prise de décision [Schiefer et al., 2002]. A notre avis, ce problème a de multiples causes :

- la faible prise en compte voire l'occultation de la phase d'analyse des besoins : la majorité des méthodes de développement de SAD passe directement aux étapes de modélisation [Mazon et al., 2005a],
- le manque de prise en compte complémentaire des besoins de tous utilisateurs et des sources de données,
- l'absence d'une méthode permettant de concevoir et de développer un système décisionnel complet (reposant sur une architecture variée comprenant différents espaces de stockage des données),
- le manque d'outils de modélisation des aspects statiques (données du système décisionnel) et dynamiques (traitements de dérivation et de préparation des données),
- le manque de mécanisme permettant de spécifier précisément les différentes étapes du processus de conception de SAD et leur réutilisation, même si plusieurs travaux ont mis en évidence que le développement de SAD repose sur la récurrence des tâches [Srivastava & Chen, 1999 ; Sen & Sinha, 2005] et requiert d'importantes ressources en temps et en hommes.

Afin de répondre à ces différents manques, nous souhaitons proposer une méthode couvrant le cycle de vie complet de développement (analyse, conception et implantation) reposant sur les critères suivants :

- cette méthode doit répondre aux besoins des décideurs tout en s'assurant que l'organisation possède les données sources nécessaires à leur élaboration. Cette méthode doit donc reposer sur une démarche mixte de façon à prendre en compte les besoins des décideurs tout en analysant les données sources ;
- l'analyse des besoins des sources et des décideurs ne doit pas être occultée et doit reposer sur un formalisme adapté. De plus, un système décisionnel étant profondément impliqué dans la stratégie d'une entreprise [Franco & De Lignerolles, 2000)], le développement d'un SAD doit prendre en compte les besoins des décideurs et notamment les besoins de pilotage ;
- les phases d'analyse et de conception ne doit pas seulement se centrer sur les données mais également sur les traitements de dérivation et de préparation de celles-ci ;
- cette méthode doit proposer un stockage des données dans des architectures hétérogènes et variées contenant différents modules ;
- la démarche doit être clairement explicitée afin de faciliter et de fiabiliser la tâche des concepteurs décisionnels. De plus, la capitalisation de la connaissance en termes de démarche et produit doit être encouragée afin de faciliter la réutilisation ;
- cette démarche de conception doit être validée dans un cadre industriel.

3.2 DEMARCHE DU TRIDENT DECISIONNEL

Pour atteindre les objectifs précédemment cités, nous proposons une méthode reposant sur les trois caractéristiques suivantes :

- la phase d'analyse intègre les besoins des décideurs et les sources de données. Les besoins des décideurs pourront être de différents niveaux (stratégiques et/ou fonctionnels). Cette analyse traite aussi bien des données que les traitements liés ;
- la phase de conception doit aider le concepteur décisionnel dans le choix de l'architecture et proposer les modèles de données adéquats pour chacun des modules de cette architecture ;
- afin d'accélérer et fiabiliser le développement de SAD, nous souhaitons paralléliser l'analyse des besoins utilisateurs et des sources, automatiser certaines tâches d'analyse et le choix d'architecture du SAD, capitaliser l'expérience et l'expertise par la définition d'un catalogue de patrons processus et enfin capitaliser l'information technique dans un catalogue de composant produit.

L'objectif de cette section est de présenter l'enchaînement des étapes de la démarche proposée.

En préambule, nous avons proposé une typologie des besoins basée sur 4 types. Les besoins **analytiques** sont liés aux analyses de données, à la qualité des données et aux processus de consolidation, d'historisation, d'archivage et de rafraîchissement des données. Les besoins **fonctionnel**s sont relatifs aux trois principales tâches de la chaîne décisionnelle, soit l'alimentation (règles de gestion,...), le stockage (volume et format des données) et la présentation (restitution et diffusion). Les besoins **non fonctionnels** portent sur la sécurité des données ainsi que sur les performances d'interrogation et de restitution de l'information. Les besoins **stratégiques** précisent les différentes politiques de l'organisation.

Ces besoins sont exprimés par trois groupes d'acteurs :

- **tactique** : besoins exprimés par les décideurs d'une composante fonctionnelle d'une organisation,
- **stratégique** : besoins exprimant les orientations stratégiques dans lesquelles doivent s'insérer le SAD ou les besoins transversaux à différentes composantes fonctionnelles exprimés par la direction (vision globale et synthétique d'une organisation),
- **système** : besoins liés aux sources de données existantes et aux équipements décisionnels actuellement déployés.

Afin de prendre en compte ces différents besoins, nous avons étendu le diagramme en Y [Roques, 2003 ; Roques & Vallee, 2004], aussi appelé 2TUP : "Two Track Unified Process" par l'ajout de la branche "Stratégique".

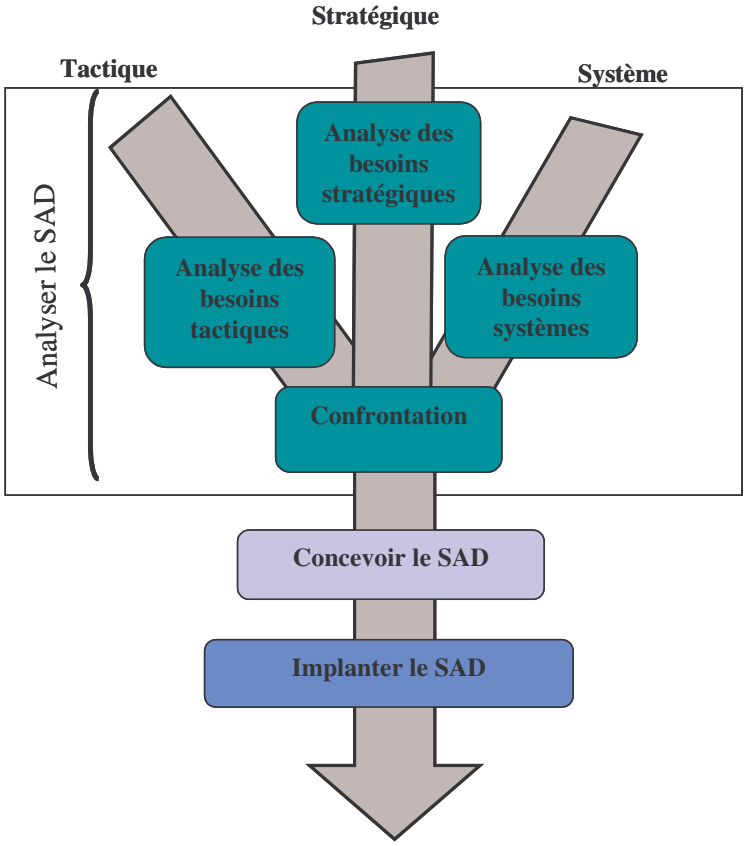


Figure 52 : Etapes du trident décisionnel

Cette extension, que nous appelons "le trident décisionnel" est composé de trois branches correspondant aux trois types de besoins énoncés précédemment. Ces trois branches sont liées à la première phase de la démarche, soit la phase "Analyser le SAD" de notre démarche. Cette phase commence par la caractérisation des groupes d'acteurs, se poursuit par l'évaluation de l'analyse des trois types de besoins et se termine par les confrontations de ces différents besoins.

La deuxième phase «Concevoir le SAD», est représentée par la partie supérieure du manche du trident décisionnel. Au cours de celle-ci, il convient de choisir l'architecture adaptée et de procéder à une conception détaillée des différents modules décisionnels de l'architecture. L'architecture adaptée et les traitements à effectuer sur les différentes sources sont définis à partir de la cartographie des besoins des acteurs et de la cartographie des sources de données.

La base du manche du trident correspond à l'implantation. Cette phase permet de définir les structures physiques des schémas de données définis à la phase précédente ainsi que les processus d'alimentation et de rafraîchissement des données.

La figure suivante donne une représentation détaillée du trident décisionnel :

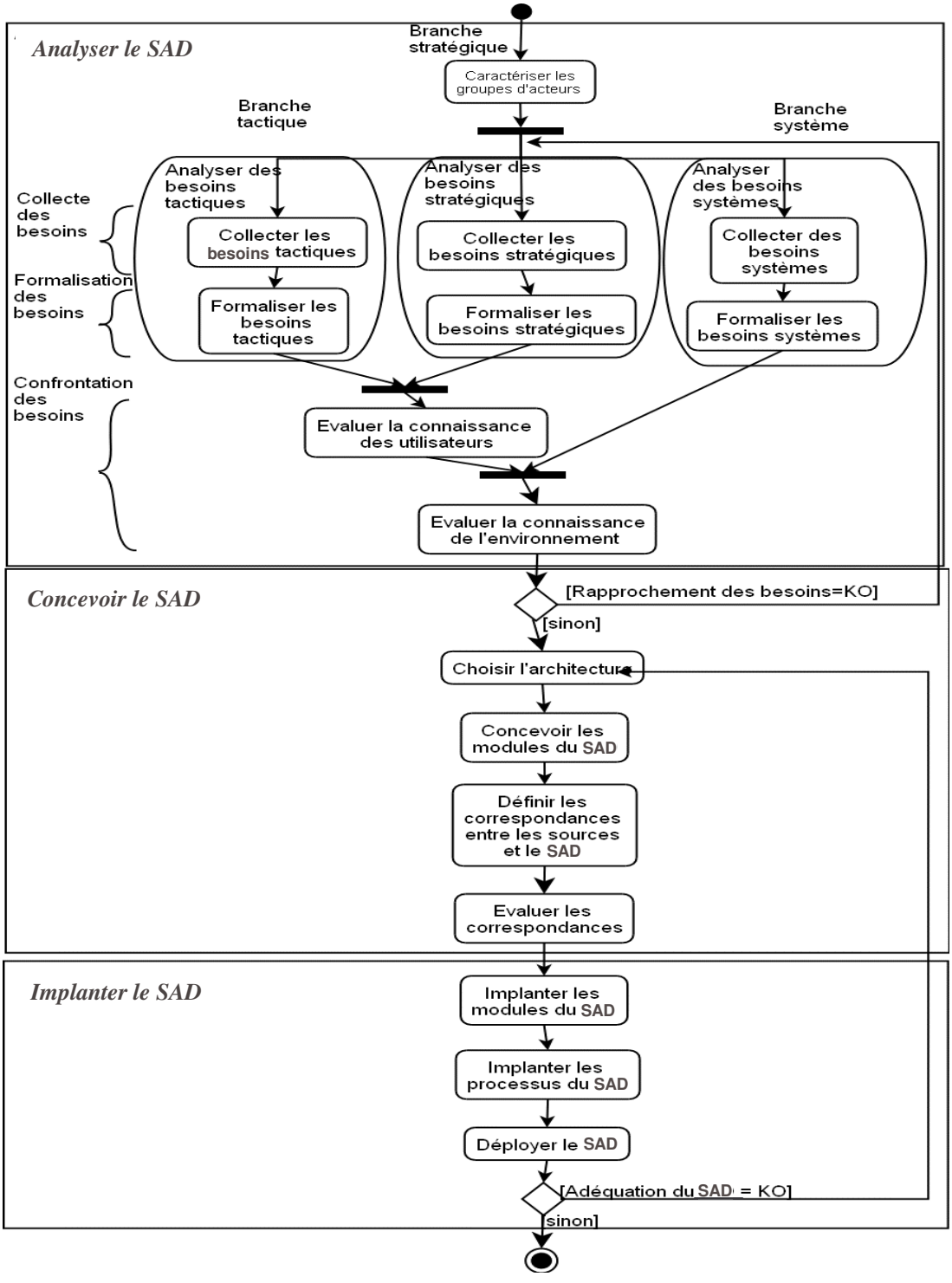


Figure 53 : Le trident décisionnel

Notre démarche repose sur un cycle itératif incrémental contenant deux points de décision : un après la tâche décisive de confrontation et un autre après le déploiement. Ce premier point permet de réitérer la phase d'analyse autant de fois que nécessaire afin de synchroniser les besoins des décideurs avec les sources avant de concevoir le SAD. Le second point permet de s'assurer que le logiciel correspond aux besoins.

3.3 ANALYSER LE SAD

L'analyse regroupe différentes étapes permettant de caractériser les acteurs et les trois types de besoins énoncés précédemment. Chacune de ces étapes est étudiée dans les sections suivantes.

3.3.1 Caractériser les groupes d'acteurs

Cette étape donne une vision globale de l'organisation en la découpant en domaines et en processus métier associés. Cette vision macroscopique est complétée par une représentation microscopique adaptée au projet. Notamment, cette étape permet de préciser le ou les domaine(s) voire le ou les sous-domaine(s) concernés par le SAD. Une fois que ce domaine est délimité, cette étape permet de recenser les acteurs types concernés par le SAD. Nous n'avons pas proposé de formalisme particulier pour cette étape ; nous proposons d'utiliser par exemple des modèles de buts [Prakash & Gosain, 2003 ; Giorgini et al., 2005 ; Mazon et al., 2005a ; Gam & Salinesi, 2006].

3.3.2 Analyser les besoins tactiques

Cette branche du trident permet d'effectuer une analyse approfondie des besoins métier des décideurs. Cette étude des besoins repose sur des interviews, des questionnaires et les documents de l'entreprise (cahier des charges, ancien rapport d'analyse…).

L'étape "Collecter les besoins tactiques" permet de préciser le type d'outil utilisé (outil de reporting, outil d'analyse décisionnelle, BD relationnelles, classeur d'un tableur…) ainsi que les données et les traitements associés. Pour la partie statique, nous proposons d'étendre l'étape de collecte des données de la démarche descendante présentée en section 2.3 afin de pouvoir analyser les aspects statiques et dynamiques.

Pour les aspects statiques, nous proposons soit d'exprimer les besoins analytiques des décideurs au travers de tableaux éventuellement croisés (formalisme couramment utilisé par les décideurs [Fernandez, 2003]), soit aux travers des mécanismes proposés dans les sections 2.3.1 (requêtes types, questionnaires et règles de gestion) et 2.3.2 (matrice des besoins).

Pour les aspects dynamiques, notre collaboration avec la société ID-6 (thèse CIFRE de Estella Annoni), nous a permis de mettre en évidence une trentaine de traitements types différents. Afin de faciliter la tâche du concepteur, nous avons proposé un formalisme graphique pour caractériser les propriétés de ces différents traitements. Ce graphe, que nous appelons graphe des propriétés se décompose en deux sous-graphes :

- le premier intitulé "décision" regroupe les propriétés en 7 catégories caractérisant les traitements liés à la préparation des données. Par exemple, la catégorie "validité" permet de préciser les caractéristiques de la conservation des données à court, moyen et long termes (rafraîchissement, historisation et archivage). La propriété "Rafraîchissement" précise la fréquence, les conditions et le mode de rafraîchissement des données décisionnelles. De même, la propriété "historisation" permet de préciser la période et la condition de l'historisation des données ;
- le second sous-graphe intitulé "technique" regroupe les propriétés en 4 catégories caractérisant les traitements ETL effectués sur les sources. Enfin, les feuilles de ce graphe correspondent aux propriétés caractérisant les traitements.

La figure suivante donne la représentation de base d'un graphe des propriétés

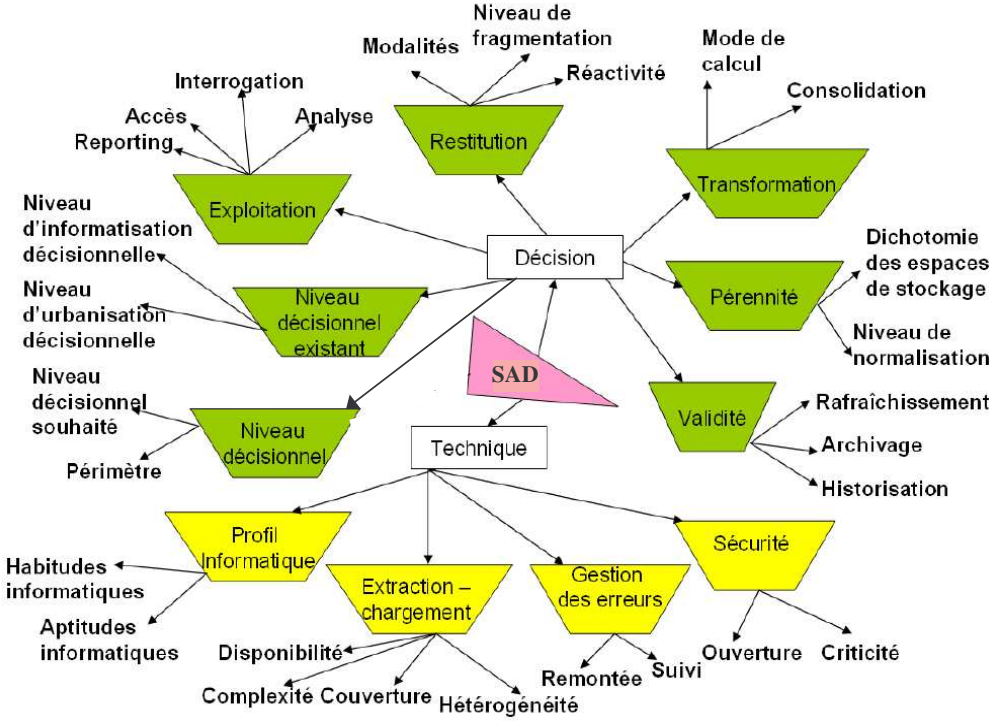


Figure 54 : Graphe des propriétés

La figure suivante donne un exemple d'instanciation de ce graphe pour un cas particulier.

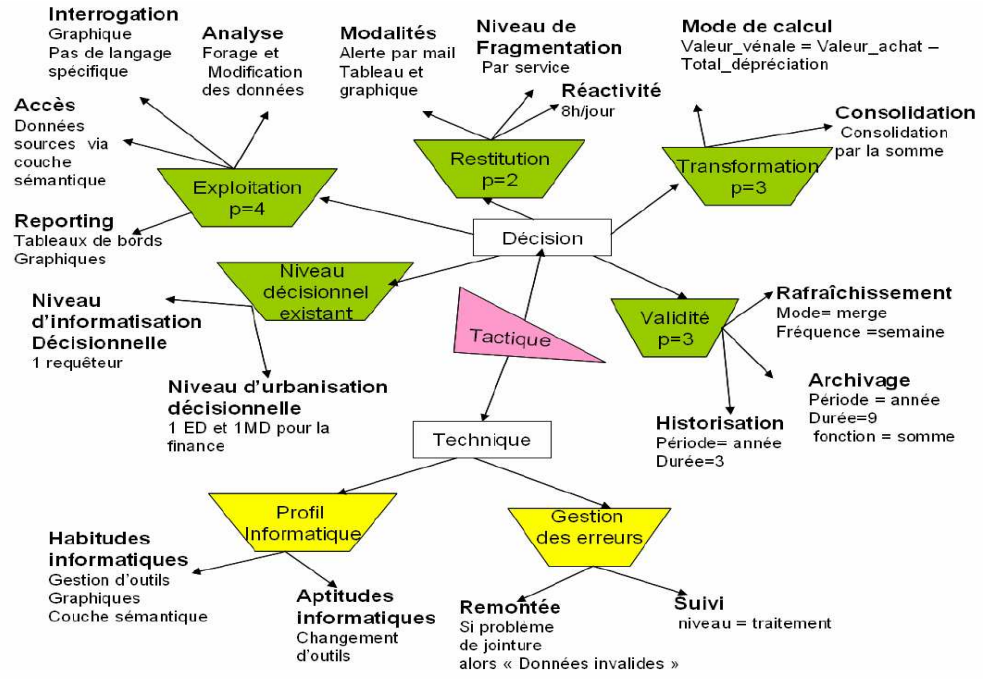


Figure 55 : Instanciation d'un graphe des propriétés

L'étape "Formaliser les besoins tactiques" est un ensemble de règles pour traduire les différents besoins exprimés précédemment à l'aide d'une représentation multidimensionnelle. Cette étape, explicitée dans [Annoni, 2007], produit un **diagramme décisionnel**. Un tel diagramme est une extension du formalisme UML permettant d'unifier dans un même schéma les aspects statiques et dynamiques [Annoni et al., 2006c ; Annoni et al., 2006d].

3.3.3 Analyser les besoins stratégiques

Ces besoins visent à traduire les orientations politiques de la stratégie d'entreprise en définissant le cadre général du développement d'un SAD en termes d'objectifs et de contraintes. Notamment, ils permettent de définir les contraintes et objectifs globaux :

- une planification globale du développement du SAD (délai, chronologie et synchronisation des développements des différents composants),
- la politique matérielle et logicielle,
- les cadres budgétaires.

Ces besoins permettent également de spécifier un ensemble de contraintes et objectifs locaux au projet. Ils permettent notamment de recenser les indicateurs de performances voire certains paramètres d'analyse utilisés dans le cadre d'une stratégie d'entreprise. Si ces besoins peuvent s'exprimer sous forme de tableaux de synthèse, nous recommandons de suivre la démarche définie dans la section précédente (3.3.2) [Annoni et al., 2006a]. Dans le cas contraire, il est possible d'utiliser les modèles de buts [Prakash & Gosain, 2003 ; Giorgini et al., 2005 ; Mazon et al., 2005a ; Gam & Salinesi, 2006]. Cette branche du trident décisionnel nécessite de nouveaux approfondissements.

3.3.4 Analyser les besoins systèmes

L'analyse des besoins systèmes permet d'étudier les sources des applications transactionnelles pour construire le diagramme décisionnel système.

L'étape de collecte permet d'obtenir le schéma conceptuel des sources pertinentes pour l'élaboration du SAD. Ce schéma conceptuel est soit directement repris des documents de l'organisation étudiée, soit construit à partir des principes de rétro-conception (reverse-engineering). Cette étape peut éventuellement être complétée par des interviews permettant de construire des graphes de propriétés.

L'étape de formalisation permet de construire le diagramme décisionnel système par application de principes similaires à ceux proposés en 2.4 (démarche ascendante). Tous les détails de cette étape sont dans [Annoni, 2007] et [Annoni et al., 2007].

3.3.5 Les confrontations

La tâche "évaluer la connaissance des utilisateurs" permet de confronter les diagrammes décisionnels des groupes tactiques et stratégiques. Notamment, cette étape permet de vérifier l'adéquation des paramètres du niveau stratégique avec ceux du niveau tactique. La tâche "évaluer la connaissance de l'environnement" permet de s'assurer que l'ensemble des données sources permet de répondre aux besoins de tous les utilisateurs. Afin de faciliter ces différentes tâches de confrontation, nous avons proposé un ensemble de règles de fusion et de confrontation de schémas. Vous trouverez tous les détails dans [Annoni, 2007].

3.4 CONCEVOIR LE SAD

Dans notre démarche (cf. Figure 53), la phase de conception regroupe les tâches "choisir l'architecture" et "concevoir les modules du SAD" (comme préconisé dans [Srivastava & Chen, 1999]) complétées par les tâches "Définir les schémas de correspondances" et "Evaluer les schémas de correspondances" afin de traduire les processus de construction et d'alimentation des SAD [Vassiliadis et al. 2002].

Le problème majeur à résoudre est que les méthodes actuelles de développement ne couvrent même pas les deux premières tâches définies par [Srivastava & Chen, 1999]. Le plus

souvent, ces méthodes se centrent uniquement sur un module décisionnel et elles ne guident pas le concepteur dans le choix d'une architecture globale.

Aussi, le point clé pour la réussite de la phase de conception est de répondre de manière fiable à la première tâche, à savoir "choisir l'architecture". Cette tâche est vitale car elle permet d'avoir une vision globale de l'architecture technique à mettre en œuvre. Pour répondre à cette problématique, nous avons proposé un algorithme [Annoni et al., 2006b].

3.4.1 Architecture modulaire

L'étude des différents projets au sein de la société I-D6, nous a permis de mettre en évidence 4 types de modules dans l'architecture d'un SAD [Annoni et al., 2005a]. En plus des ED et MD, nous trouvons :

- Passerelle Décisionnelle (PD) : module d'accès direct aux sources afin de présenter les données aux décideurs ;
- Système de Collecte des Informations (SCI) ou source globale : système intermédiaire qui permet de collecter l'information en amont de la chaîne décisionnelle. Cet espace de stockage temporaire permet la transformation et la consolidation des données issues de sources réparties, mobiles et hétérogènes. Il offre ainsi une manipulation plus simple et centralisée des sources de données. Cet espace de stockage permet également de réduire la charge du système opérationnel due aux alimentations du SAD.

La figure suivante illustre ces différents modules architecturaux :

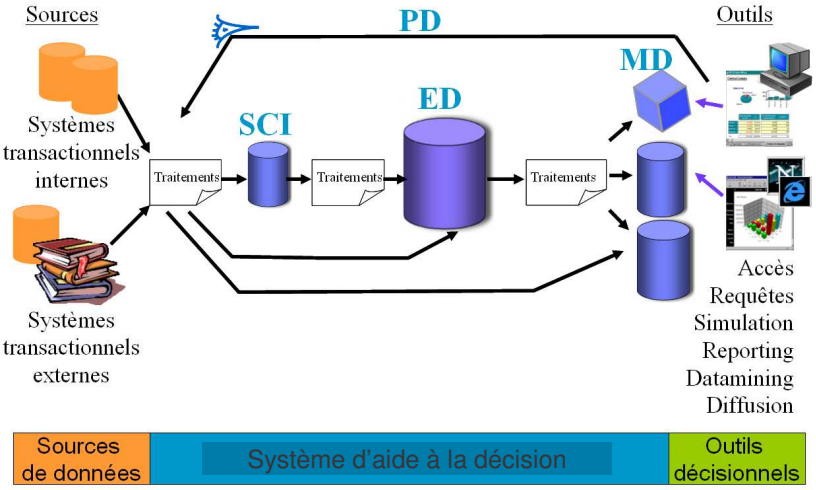


Figure 56 : Les différents modules d'un système décisionnel

Les architectures décisionnelles peuvent être composées de 1 à 4 modules architecturaux différents. L'architecture est vide quand l'organisme n'a pas de système d'information décisionnel. Dans le cas contraire, elle doit respecter les règles de cohérence suivantes :

- si l'architecture est composée de plus d'un magasin, alors il existe nécessairement un entrepôt de données pour la centralisation de l'information,
- l'architecture peut être constituée de plusieurs modules du même type, mais au plus d'un seul de type entrepôt. Ceci permet d'avoir toute l'information centralisée en un seul espace et facilite l'interrogation des données historisées,
- un SCI est un système de stockage temporaire et un PD est un accès direct aux sources donc un système décisionnel ne peut être constitué uniquement de {SCI} ou de {PD, SCI}.

Dans le tableau suivant, nous recensons les 13 architectures décisionnelles possibles en fonction des règles énoncées précédemment.

	Nombre de type de modules			
Architecture	**1**	**2**	**3**	**4**
	{PD$^{1..N}$}	{PD$^{1..N}$,MD$^{1..N}$}	{PD$^{1..N}$,MD$^{1..N}$, ED}	{PD$^{1..N}$,SCI$^{1..N}$,MD$^{1..N}$,ED}
	{MD$^{1..N}$}	{PD$^{1..N}$,ED}	{PD$^{1..N}$,SCI$^{1..N}$,MD$^{1..N}$}	
	{ED}	{SCI$^{1..N}$,MD$^{1..N}$}	{PD$^{1..N}$,SCI$^{1..N}$,ED}	
		{SCI$^{1..N}$,ED}	{SCI$^{1..N}$,MD$^{1..N}$,ED}	
		{MD$^{1..N}$,ED}		

Figure 57 : Architectures possibles

Pour faciliter la tâche du concepteur, nous avons défini une fonction permettant de déterminer le nombre et le type de modules de l'architecture [Annoni et al., 2006b]. Cette fonction tient compte des 5 arguments présentés ci-après :

- le niveau de couverture des données (NCD) : la couverture d'un SAD est soit verticale s'il se rapporte à un seul métier ou une seule classe d'utilisateurs, et transversal sinon ;
- le niveau de traitement des données (NTD) : ce niveau précise la qualité des traitements à effectuer sur les données sources pour satisfaire les besoins décisionnels. Les traitements sont qualifiés de "beaucoup travaillés" si les traitements contiennent de nombreuses agrégations ou consolidations, et "peu travaillé" dans le cas contraire ;
- le niveau d'équipement décisionnel existant (NEE) : ce niveau précise l'architecture décisionnelle existante au sein de l'organisme. Ce niveau prend la valeur de l'une des 13 combinaisons valides de notre typologie définie précédemment ;
- le niveau de complexité des sources (NCS) : ce niveau binaire ("peu complexe" ou "complexe") permet d'évaluer les difficultés et les exigences d'interrogation des sources de données. Une source est qualifiée de complexe s'il faut consolider et centraliser ses données afin de réduire la charge du système ;
- le niveau décisionnel souhaité (NDS) : ce niveau binaire ("partiel", "complet") précise la complétude de l'architecture décisionnelle souhaitée.

Cette fonction retourne le nombre des différents types de modules. Des exemples d'application de la fonction sont donnés dans [Annoni et al., 2005a ; Annoni et al., 2006b].

3.4.2 Concevoir les modules

Cette tâche consiste à fournir les schémas des différents modules.

Un SCI, recensant les données sources pour la plupart stockées dans des SGBD relationnels, repose sur une base de données relationnelles. Sa conception repose donc sur les principes classiques de conception de bases de données relationnelles (schéma entité-association ou diagramme de classes traduit en un schéma relationnel).

Pour la conception d'un entrepôt de données, il suffit de construire un diagramme de classes UML étendu tel que nous l'avons proposé dans le second chapitre de ce mémoire.

La conception des magasins de données dépend de son type. Si le MD repose sur un outil de manipulation multidimensionnelle, il faut concevoir un schéma multidimensionnel comme présenté dans la première section de ce chapitre. Si le MD repose sur une BD relationnelle, il suffit d'utiliser les principes de la conception d'une telle base. Si le MD repose sur un outil de reporting, il faut soit définir la BD servant de support à l'élaboration des rapports soit directement définir la structure des rapports. Si le MD repose sur un classeur d'un tableur, il suffit

de reprendre les techniques de conception d'un classeur précisant le contenu des différentes feuilles de celui-ci.

3.5 PRINCIPES DE REUTILISATION

Dans le cadre de notre collaboration avec la société I-D6, nous avons constaté de nombreux points communs dans les processus de développement de projets décisionnels. Ce constat est corroboré par les travaux de [Sen & Sinha, 2005]. A chaque début de projet, il n'est pas rare que des tâches du processus d'ingénierie soient réutilisées de façon empirique par les concepteurs décisionnels. Cette réutilisation est opportuniste car elle dépend de l'expérience, de la capacité de corrélation et de la politique de réutilisation des concepteurs affectés à ce projet. De plus, ce type de réutilisation est généralement non documenté et non disponible pour tous. Cette re-définition quasi systématique du processus représente "une perte de temps" ainsi qu'un coût pour les organisations.

Ainsi, face à ces nombreux points communs dans les processus de développement des SAD d'architectures variées, nous souhaitons capitaliser la connaissance et l'expertise des concepteurs décisionnels. Nous souhaitons proposer un catalogue de composants réutilisables pour l'analyse et la conception de SAD et de faciliter leur réutilisation.

Les travaux relatifs à la capitalisation et à la réutilisation dans le domaine des SAD sont à leurs débuts. Seules, deux propositions fournissent des patrons produits d'analyse [Jones & Song, 2005] ou de conception logique en étoile [Feki et al., 2006]. Cependant, ces propositions sont limitées à des domaines d'activité prédéfinis [Feki et al., 2006], à la définition d'un seul composant comme les dimensions [Jones & Song, 2005] et ne concernent qu'une partie d'une phase du processus d'ingénierie. De plus, elles ne facilitent pas et elles ne systématisent pas la réutilisation car les concepteurs ne sont pas guidés pour la mise en place du SAD via des patrons processus. Ces patrons sont des blocs indépendants où les relations entre les patrons ne sont pas définies. Les auteurs ne proposent pas d'outil pour concevoir un SAD par réutilisation.

3.5.1 Catalogue de patrons

Notre objectif est de systématiser la réutilisation lors des projets décisionnels, faciliter la réutilisation aussi bien par les concepteurs débutants qu'experts, améliorer la communication et la traçabilité de la documentation. En effet, comme le soulignent les travaux de [Barbier et al. 2004], la réutilisation de composants permet de réduire les coûts et les délais de conception, d'implantation et de maintenance si elle s'allie à la traçabilité.

Pour répondre à nos besoins, nous proposons un catalogue de patrons d'analyse, de conception et produit [Annoni et al., 2005a]. Le concept de patron a été choisi car il s'agit d'un composant dont le contenu est disponible et adaptable par les concepteurs décisionnels, soit des boîtes blanches, avec un degré de variabilité qui favorise une adaptation simple. Les patrons d'analyse et de conception de notre catalogue sont définis suivant l'approche orientée problème des composants par des triplets <Problème, Solution, Contexte> [Alexander, 1977; Gamma et al., 1995]. Le problème est un élément de démarche à réaliser ou un produit à définir. La solution est une démarche ou un modèle représenté respectivement par des algorithmes, des diagrammes d'activités, des modèles de documents ou des modèles de données. Enfin, le contexte définit la situation pour laquelle la solution décrite dans le composant est utilisable.

Notamment, nous souhaitons proposer un formalisme permettant l'expression du contexte du patron et générant la documentation projet [Annoni et al., 2005a ; Annoni et al., 2005b].

Le catalogue guide le concepteur décisionnel via les relations inter-patrons existantes [Rieu, 1999] durant les deux premières phases de l'ingénierie des SAD, ce qui facilite et contribue à systématiser la réutilisation des patrons. Nous proposons d'utiliser un formalisme structuré, et

plus précisément d'étendre le formalisme P-SIGMA [Conte et al., 2001] pour formaliser l'expression du contexte du patron et générer la documentation projet.

Notre choix s'est porté sur P-SIGMA [Conte et al., 2001], car ce formalisme est une tentative d'unification des formalismes structurés qui ont été proposés [Coad, 1992; Gamma et al. 1995; Buschmann et al., 1996]. Notamment P-SIGMA présente les avantages suivants : il intègre l'expression des aspects produits et processus, il propose un grand nombre de relations inter-patrons et il facilite la sélection, la réutilisation et l'organisation des composants. Ce dernier avantage est basé sur le regroupement de l'information en trois parties :

- interface : elle permet la sélection du patron,
- réalisation : elle exprime la solution proposée par le patron,
- relation : elle permet d'organiser les relations entre patrons.

L'extension proposée intervient à deux niveaux :

- Intégration de la documentation : avec deux rubriques "Documents sources" précisant la documentation utilisée au cours de la solution démarche pour obtenir la solution modèle et "Documents cibles" correspondant à la documentation résultant de l'application de la solution démarche ;
- Formalisation du contexte précisant les situations dans lesquelles un problème est résolu. Pour ce faire, nous avons proposé un formalisme dont vous trouverez tous les détails dans [Annoni, 2007].

Afin de faciliter l'utilisation des patrons, nous avons proposé un catalogue contenant quatre niveaux d'éléments processus reposant sur la démarche du Trident décisionnel. L'élément de plus haute granularité est l'étape. Dans notre catalogue, nous avons proposé les étapes "mettre en place le SAD" et "maintenir le SAD". Une étape se décompose en phases qui elles-même se décomposent en tâches voire en sous-tâches. Les patrons produits reposent sur des modèles de documents, des modèles de tableaux, des modèles de graphes ou des modèles conceptuels de données et de traitements.

La Figure 46 présente le diagramme de collaboration de patrons [Fowler, 2004] de l'étape "Mettre en place un SID". Ce diagramme définit les relations entre les principaux patrons relatifs aux phases, tâches, et sous-tâches de l'étape d'analyse.

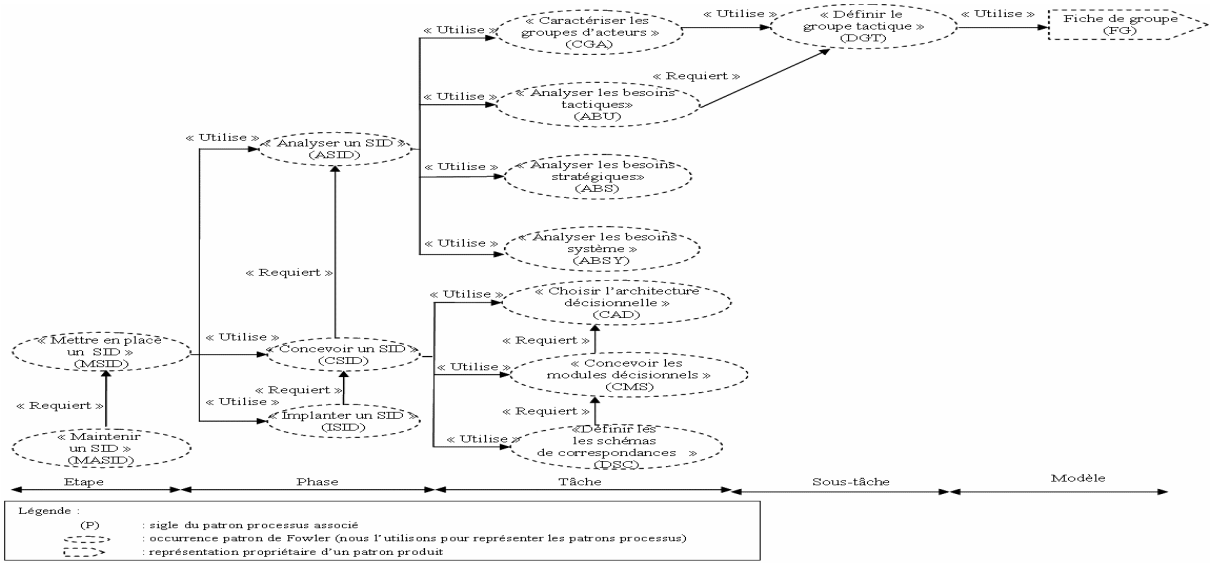


Figure 58 : Catalogue de patrons

La figure suivante présente le formalisme P-SIGMA étendu sous forme tabulaire que nous utilisons pour la définition de notre catalogue.

Parties	*Rubriques*	*Champs*
INTERFACE	Sigle	ASID
	Identifiant	Analyser un SID
	Classification	analyse ∧ processus ∧ SID
	Contexte	$\forall x \in N_+^*, \forall y \in N_+^*, (x \leq k, y \leq k \Rightarrow (u_s(x,y) = 1, r_s(MSID, x, y) = 1))$ avec x : itération en cours du processus de développement du SID, y : nombre d'utilisations de ce patron au cours du processus complet k : nombre maximal d'itérations de la phase d'analyse u_s : la fonction qui indique les conditions dans lesquelles ce patron est utilisable r_s : la fonction qui indique les conditions dans lesquelles ce patron requiert le patron MSID
	Problème	Comment analyser un système d'information décisionnel ? L'analyse du SID correspond à l'étude de la mise en place du SID. Elle est réalisée afin d'évaluer la faisabilité du projet, en l'occurrence le rapprochement entre les besoins
	Motivation	L'analyse du SID consiste à caractériser les groupes d'acteurs et analyser leurs besoins en vue de les rapprocher.
	Forces	Ce patron guide le concepteur pour l'analyse d'un système d'information décisionnel. Il définit le contexte d'exploitation du SID
REALISATION	Solution-démarche	La solution consiste à la mise en oeuvre de ce diagramme d'activités. 
	Solution-modèle	
	Cas d'applications	Au début d'un projet décisionnel ou quand les besoins n'ont pas été rapprochés ou lorsque que les besoins ont été modifiés
	Conséquences d'application	Définition des acteurs et de leurs besoins
	Documents sources	Le cahier des charges
	Documents cibles	
RELATION	Utilise	« Caractériser les groupes acteurs », « Analyser les besoins tactiques », «Analyser les besoins stratégiques », « Analyser les besoins systèmes », « Evaluer la connaissance »
	Requiert	«Mettre en place un SID »
	Raffine	
	Alternative	

Figure 59 : Patron "Analyser le système d'information décisionnel"

3.6 IMPLANTATION AU SEIN DE LA SOCIETE I-D6

Ces travaux ont été menés dans le cadre d'une collaboration avec la société de services informatiques I-D6, spécialisée dans le domaine du décisionnel. Cette collaboration nous a permis de valider nos propositions en terme de processus. Pour faciliter l'acception de cette méthode au sein de la société I-D6, nous avons suivi les recommandations de la démarche qualité "Roue de Deming" [Annoni et al., 2006e]. Cette démarche qualité repose sur un cycle de vie itératif incrémental comme modélisé dans la figure suivante :

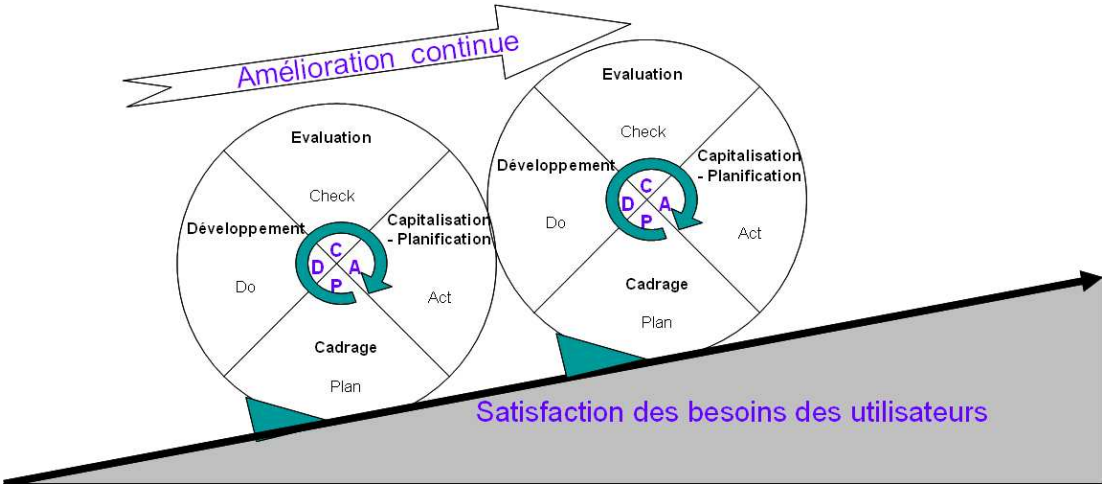


Figure 60 : Démarche qualité Roue de Deming

Pour répondre aux besoins de la société I-D6, nous avons itéré quatre fois la Roue de Deming. Dans un premier temps, nous avons étudié la documentation existante des projets réalisés au sein de la société afin de dresser un état de l'existant des modèles, des méthodes, des techniques, des outils utilisés par I-D6. Cet état a mis en avant un besoin d'homogénéisation et d'harmonisation des terminologies et des méthodes utilisées au sein de la société. Nous avons donc regroupé les concepteurs décisionnels en trois groupes suivant l'expertise dans l'ingénierie des SAD (expert, intermédiaire, débutant). A chaque itération, nous avons ajouté un nouveau groupe de collaborateurs : la première itération a été faite avec les experts et la dernière avec les débutants. A chaque étape, nous avons posé nos hypothèses, nos bases et nos objectifs afin de mettre en avant des éléments de solution. Chaque itération a été clôturée par la remise de fiches d'évaluation. Toutes les itérations n'ont pas eu la même durée comme présenté dans la figure suivante :

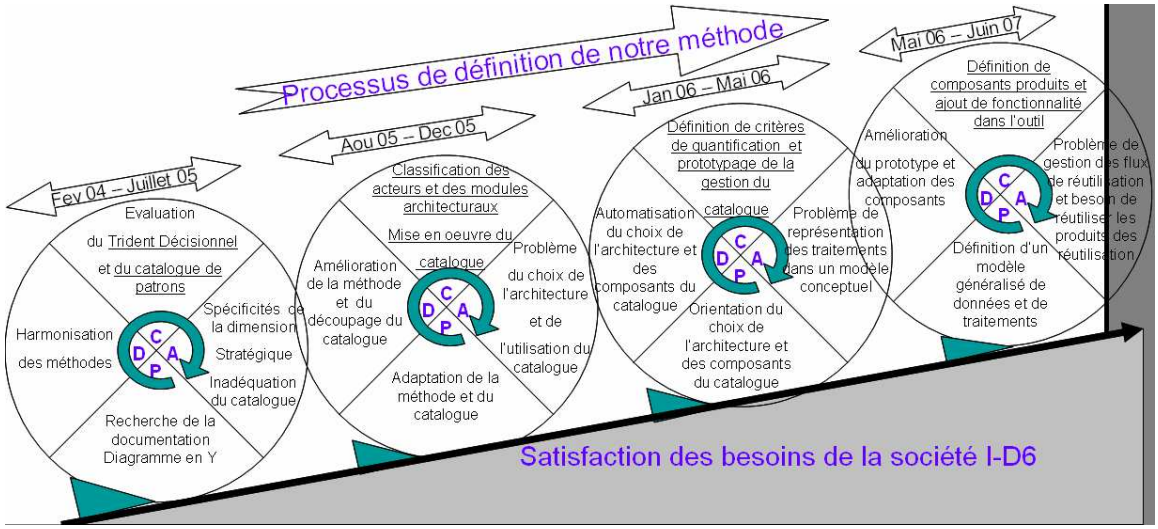


Figure 61 : Itérations de la roue de Deming

L'application démarche qualité nous a permis de valider nos propositions et d'effectuer un certain nombre de corrections. Notamment, nous avons affiné la démarche du trident et nous avons dû revoir plusieurs fois la définition des patrons. Enfin, cette intégration industrielle, nous a permis de mettre en avant le besoin d'un outil d'aide à la conception reposant sur notre démarche du trident et les différents patrons de conception. Cette première version de l'outil doit être encore améliorée pour une adhésion plus forte des collaborateurs de la société I-D6.

4 BILAN ET PERSPECTIVES

Ce chapitre vise à compléter les propositions effectuées dans les chapitres précédents (modèles d'entrepôt et de magasins) en se focalisant sur la proposition d'une démarche de conception d'un système d'aide à la décision.

A notre connaissance, il n'y a pas à l'heure actuelle de proposition relative à une démarche de conception de système décisionnel complet (intégrant différents modules tels que les entrepôts et les magasins de données).

4.1 CONCEPTION D'UN MAGASIN DE DONNEES MULTIDIMENSIONNELLES

Dans un premier temps, nous nous sommes centrés sur le composant spécifique d'un SAD, à savoir les magasins de données multidimensionnelles. Trois types de démarche de conception ont été proposés. Les ***méthodes ascendantes*** partent des sources de données pour concevoir les schémas multidimensionnels et présentent l'inconvénient de ne pas prendre en compte les besoins des décideurs. Les ***méthodes descendantes*** prennent en compte les besoins des décideurs au risque de déboucher sur des schémas non implantables par manque de données. Les ***méthodes mixtes*** combinent les deux précédentes. Comme pour les méthodes descendantes, les différentes propositions manquent de précision quant à la collecte des besoins utilisateurs et la transformation de ces besoins en un schéma conceptuel. De plus, à l'exception des travaux de [Bonifati et al., 2001], ces méthodes prennent en compte le contenu des données sources mais ne construisent pas réellement de schémas multidimensionnels à partir des sources.

Afin de compléter notre proposition du chapitre 3, nous avons proposé une démarche pour la conception de **bases de données multidimensionnelles contraintes**. Ces contraintes permettent d'assurer la fiabilité des données et la cohérence des analyses décisionnelles. La démarche que nous avons proposée repose sur une **approche mixte** car elle permet de tenir compte des besoins des décideurs et de la description des données sources. Cette démarche mixte repose sur une approche descendante, une approche ascendante et une confrontation.

Contrairement aux autres propositions, nous avons **clairement identifié les étapes à suivre** lors de l'approche descendante. Nous avons identifié trois étapes : collecte des données, spécification des besoins et formalisation. L'étape de collecte des données repose sur les principes de requêtes types, questionnaires et règles de gestion. L'étape de spécification des besoins permet de définir les matrices de besoins des attributs et le recensement de contraintes issues de la traduction des règles de gestion. L'étape de formalisation des besoins consiste en la traduction des matrices des besoins et des contraintes en un schéma multidimensionnel contraint qualifié de schéma "idéal".

L'approche ascendante ne se limite pas à une simple consultation des données sources mais en une réelle construction d'un schéma conceptuel multidimensionnel. Pour la spécification de ce schéma, nous avons défini un processus en 6 étapes. Ce processus repose sur la détection des classes représentatives et déterminantes dans l'entrepôt de données sources. Les classes représentatives permettent de définir un fait et les classes déterminantes permettent de déterminer les dimensions. Cette approche ascendante aboutit à un schéma multidimensionnel qualifié de schéma "candidat".

L'étape de confrontation permet de définir le schéma final à partir d'un schéma "idéal" et d'un schéma "candidat". Cette étape permet de spécifier un schéma multidimensionnel contenant une définition correcte de la granularité d'analyse, des données issues des besoins décisionnels ou de sources, des données cohérentes en fonction des sources et des contraintes.

Dans le cadre de mes enseignements, je propose une version simplifiée de cette démarche (non intégration des contraintes). Lors de leurs stages, les étudiants utilisent ce processus de

conception ou l'adaptent en fonction des préconisations des entreprises d'accueil. Ce retour d'expérience est intéressant et nous permet de nous assurer que cette démarche peut être utilisée dans un cadre professionnel. Notamment, par opposition aux Systèmes d'Information où nous trouvons de nombreuses solutions en terme de processus et produit, les sociétés spécialisées dans le décisionnel sont toujours à la recherche d'une méthode de conception reconnue et fiable.

4.2 CONCEPTION D'UN SYSTEME D'AIDE A LA DECISION

Dans un second temps, nous avons proposé une solution pour la conception d'un système d'aide à la décision complet. Un système d'aide à la décision étant lié à la stratégie d'une entreprise, son développement doit prendre en compte les besoins des décideurs, les données sources et les besoins de pilotage. Cette particularité a abouti à la proposition d'une démarche reposant sur un trident décisionnel. Ce trident comprend trois branches correspondant aux trois types de besoins : tactique, stratégique et système.

Pour l'analyse des besoins tactiques, nous proposons d'étendre la proposition précédente afin de modéliser les données et les traitements (d'historisation, d'archivage et de rafraîchissement) associés. Les besoins stratégiques permettent d'étudier et de délimiter le domaine d'étude tout en intégrant les contraintes de planification du développement du SAD, de politique matérielle et logicielle et de cadre budgétaire. Ces besoins stratégiques peuvent également se traduire par des diagrammes décisionnels. Les besoins systèmes permettent de dériver des schémas multidimensionnels à partir des données sources.

Le développement d'un système d'aide à la décision repose sur l'intégration de différents modules. Afin d'aider le décideur dans la spécification de ces modules, nous avons développé une fonction permettant de préciser de manière automatique le nombre de modules. Cette fonction prend en entrée 5 critères : le niveau de couverture des données (verticale ou transversale), le niveau de traitement des données (peu ou beaucoup travaillé), le niveau d'équipement décisionnel existant, le niveau de complexité des sources (peu complexe ou complexe) et le niveau décisionnel souhaité (partiel ou complet). De plus, afin de capitaliser la connaissance et l'expertise des concepteurs décisionnels afin d'être réutilisées, nous avons proposé un catalogue de patrons. Cette méthode de conception a été validée au sein de la société I-D6 au travers de la démarche qualité "Roue de Deming".

4.3 PRODUCTION SCIENTIFIQUE

Les travaux relatifs à la démarche de modélisation ont été supportés par deux thèses que j'ai co-encadrées. La première [Ghozzi, 2004], a permis de proposer une démarche pour la modélisation multidimensionnelle contrainte. La seconde [Annoni, 2007] a permis de compléter ce travail en proposant une méthode de conception de SAD reposant sur une architecture modulaire. Cette méthode identifie les phases d'analyse, de conception et de développement tout en intégrant un catalogue de patrons pour la capitalisation et la réutilisation.

Cette dernière proposition a fait l'objet d'une collaboration CIFRE avec la société de services I-D6. Afin de faciliter l'acceptation de cette méthode, Estella Annoni a appliqué la démarche qualité de la "Roue de Deming" [Annoni et al., 2006e]. Cette démarche nous a permis de faire 4 itérations avec un nombre croissant de groupes de collaborateurs afin de satisfaire leurs besoins. De plus, ces travaux ont servi de base à une nouvelle collaboration avec le laboratoire GRIMMAG[17] afin de proposer de nouvelles solutions pour l'intégration des traitements ETL dès la phase d'analyse.

[17] Tous les détails de cette collaboration sont exposés dans le chapitre 6.

Ces travaux ont également donné lieu à l'encadrement de deux Masters Recherche 2IH de l'Université Paul Sabatier. F. Boucheikh a fait un état de l'art des méthodes de conception de SAD et a proposé une première version [Boucheikh, 2005]. Quant à M. Gargouri [Gargouri, 2006], il a proposé un outil permettant de construire des schémas multidimensionnels à partir de sources relationnelles.

D'un point de vue publications, nous pouvons citer les références suivantes[18] :

- 3 articles dans des conférences internationales : SEKE'07 [Annoni et al., 2007], DAWAK'06 [Annoni et al., 2006a], DEXA'06 [Annoni et al., 2006b],
- 2 articles dans des revues nationales : RSTI-ISI [Annoni et al., 2005a], RSTI-ISI [Annoni et al., 2006c],
- 4 articles dans des conférences nationales : EDA'06 [Annoni et al., 2006d], INFORSID'06 [Annoni et al., 2006e], EDA'05 [Ghozzi et al., 2005], AIM'05 [Annoni et al., 2005b].

4.4 PERSPECTIVES

Une première version de l'outil d'aide à la conception a été développée afin d'identifier et ordonnancer les différentes étapes du processus de conception. Une perspective à court terme est le développement d'un outil complet permettant à la fois la spécification des patrons du catalogue et leurs utilisations. Cet outil comporterait deux modules. Le premier reposerait sur un environnement graphique pour la définition des différentes rubriques d'un patron. Le second permettrait au concepteur de sélectionner le patron désiré et gérerait les différents dépôts de documents associés aux phases de développement. Cet outil permettrait d'améliorer l'interface fruste de notre première version.

Une perspective à plus long terme est de compléter nos travaux relatifs à la phase d'analyse. Dans notre proposition, nous n'avons intégré qu'une partie de ces besoins stratégiques. Notamment, la modélisation orientée buts s'avère prometteuse.

[18] Le contenu de chaque article est résumé dans le chapitre 6.

CHAPITRE VI : OUTILS, PROJETS ET PUBLICATIONS

PLAN DU CHAPITRE

1 INTRODUCTION 133

2 LOGICIEL D'AIDE A LA CONCEPTION GRAPHIQUE D'ENTREPOTS ET DE MAGASIN DE DONNEES 133

2.1 Problématique 133
2.2 Principes 134
2.2.1 Description générale 134
2.2.2 Interface graphique 135
2.2.3 Stockage 135
2.2.4 Traitements 135
2.3 Elaboration graphique et incrémentale d'un entrepôt 135
2.3.1 Extraction et organisation 136
2.3.2 Historisation 137
2.3.3 Configuration 137
2.4 Elaboration de magasins multidimensionnels contraints 137

3 OUTIL DE MANIPULATIONS MULTIDIMENSIONNELLES 139

3.1 Architecture 140
3.2 Langage assertionnel 140
3.3 Langage graphique 141
3.3.1 Interface de visualisation 141
3.3.2 Interrogation graphique 143

4 CONCEPTION ET MANIPULATION D'ENTREPOTS DE DOCUMENTS 143

4.1 Architecture de DOCWARE 144
4.2 Parseur 144
4.3 Moteur OLAP 145

5 PROJETS ET COLLABORATIONS 145

5.1 Vision synthétique 146
5.2 Présentation des projets et/ou collaborations 146

6 ENCADREMENTS ET PUBLICATIONS 148

6.1 Encadrements 148
6.2 Publications 150

7 BILAN ET SYNTHESE 153

1 INTRODUCTION

Les précédents chapitres décrivent les travaux que nous avons menés dans le cadre de la conception d'entrepôt et de magasins de données ainsi que pour la manipulation multidimensionnelle des données. L'objectif de ce dernier chapitre est de décrire les différentes productions qui ont permis de mettre en œuvre et de valider les concepts et principes proposés. Nous présentons les différents prototypes que nous avons développés, les projets auxquels nous avons participé ainsi que les encadrements et publications effectués.

L'équipe SIG de l'IRIT possède une forte expertise dans le domaine des langages visuels pour la manipulation de BD relationnelles (HQL, HQL+, CHQL) et orientées objet (OHQL, VOHQL). Nous souhaitons profiter de cette expérience pour proposer des interfaces interactives et attractives dans les prototypes d'aide à la conception et à la manipulation de SAD

Le premier prototype que nous avons défini vise à aider le concepteur dans l'élaboration d'entrepôts et de magasin de données multidimensionnelles. Ce prototype doit supporter les modèles d'ED et de MD et valider la démarche ascendante d'élaboration de magasin de données multidimensionnelles (construction d'une constellation à partir d'un ED historisé). Ce prototype doit reposer sur des représentations graphiques de schémas conceptuels et un processus d'élaboration graphique des MD.

Le second prototype complète le précédent. Après la construction de magasins de données multidimensionnelles, il faut offrir un environnement graphique pour supporter les différentes manipulations multidimensionnelles décrites dans le troisième chapitre. Plus particulièrement, ce prototype permet de manipuler une constellation de données à l'aide d'un langage graphique et d'un langage assertionnel.

Le dernier prototype permet de construire et de manipuler des entrepôts de documents. Ce prototype doit permettre de valider notre processus d'alimentation d'entrepôts documentaires à partir de documents XML et doit offrir un environnement graphique adapté pour la manipulation multidimensionnelle des données contenues dans un tel entrepôt.

Les autres productions scientifiques se matérialisent au travers d'articles dans des revues et conférences internationales et nationales. Notre objectif était de valider l'ensemble des outils méthodologiques que nous avons présentés dans les chapitres précédents (modèles d'entrepôts et de magasins, démarche de conception, langage de manipulation de données multidimensionnelles). La plupart des publications ont été écrites avec des personnes que j'ai encadrées. Ce chapitre vise également à recenser ces différents encadrements.

Enfin, nous souhaitons présenter les différents projets auxquels nous avons participés. Ces projets ont abouti à des collaborations avec des laboratoires de recherche et des entreprises.

Les sections suivantes permettent de présenter de manière détaillée ces différents outils ainsi que les projets auxquels j'ai participé sans oublier les encadrements et les publications effectués.

2 LOGICIEL D'AIDE A LA CONCEPTION GRAPHIQUE D'ENTREPOTS ET DE MAGASIN DE DONNEES

2.1 PROBLEMATIQUE

Pour le développement de systèmes décisionnels, peu d'outils sont disponibles sur le marché. Les outils les plus usités se situent à un niveau d'abstraction logique ou physique et permettent d'assurer les fonctions ETL ("Extraction, Transformation Loading"). Ces outils ETL permettent d'extraire les données d'une BD ou d'un fichier source pour y appliquer des processus

de transformation des données et les charger dans une BD cible. Dans ces outils, la source et l'entrepôt ne sont pas représentés à l'aide de schémas conceptuels couramment manipulés par des concepteurs. Avec ces outils actuels, le schéma logique de l'entrepôt doit être construit au préalable ; l'administrateur a la charge d'associer chaque attribut cible de ce schéma à un attribut d'une source de données ; cette phase est complexe, fastidieuse et demande une connaissance importante des structures sources. Ce type d'outil n'est pas adapté pour un processus de conception.

Profitant de l'expérience de notre équipe, notre objectif est de proposer un outil d'aide à la conception graphique et incrémentale d'entrepôts et de magasins de données.

2.2 PRINCIPES

Notre prototype d'aide à la conception repose sur une représentation conceptuelle des schémas de la source, de l'entrepôt et des magasins de données. Cet outil permet une construction graphique et incrémentale des entrepôts et des magasins à partir d'une source globale et génère automatiquement le contenu de ED et des MD. Les représentations et les manipulations graphiques présentent les avantages suivants :

- une vision graphique plus explicite de la réalité modélisée que celle offerte par les langages textuels, tabulaires ou iconiques [Canillac 1991 ; Le Parc 1997] ;
- un mode d'expression incrémentale des opérations d'extraction puisqu'une opération est définie pas à pas en progressant dans le graphe ;
- une représentation de la portée des opérations appliquées sur les graphes.

2.2.1 Description générale

La Figure 62 décrit l'architecture générale de notre prototype. Elle comprend trois niveaux : une interface graphique directement accessible par le concepteur, un module de traitement permettant de traduire les actions graphiques en requêtes sur les différents espaces de stockage.

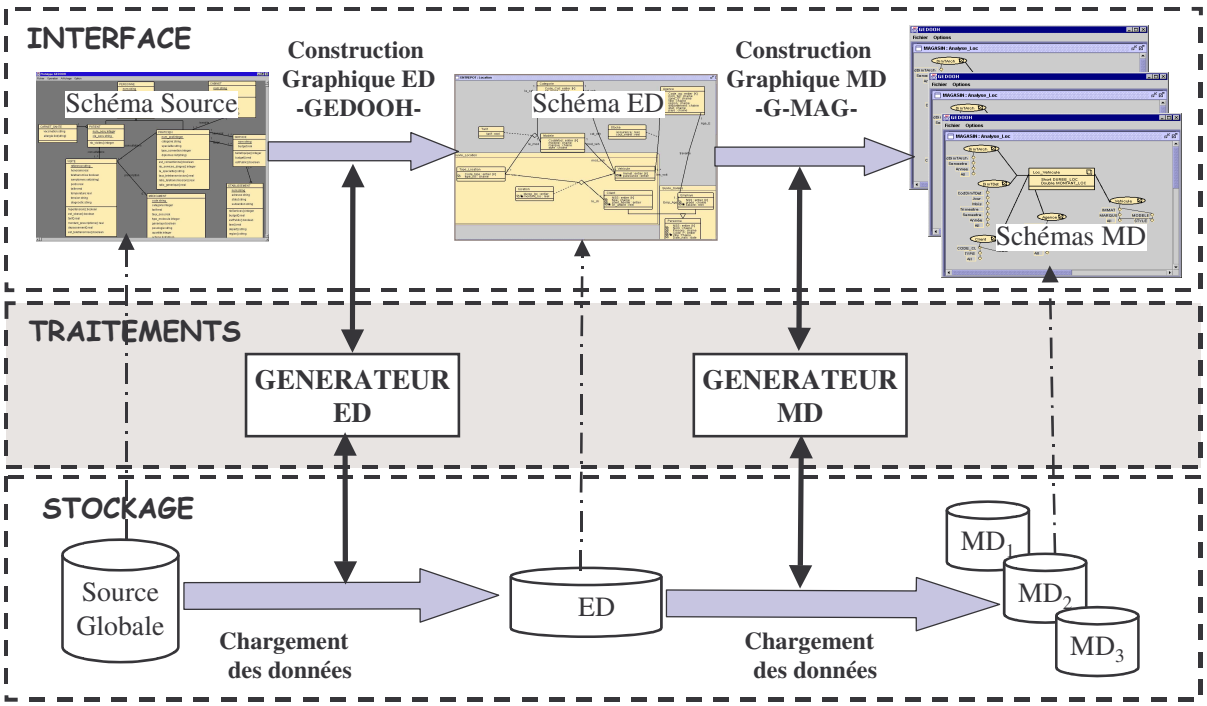


Figure 62 : Architecture du prototype de conception de système décisionnel

2.2.2 Interface graphique

L'interface graphique permet d'afficher dans différentes fenêtres la représentation conceptuelle des différents espaces de stockage associés au système décisionnel. La source globale est représentée à l'aide d'un diagramme de classes UML. L'entrepôt de données est affiché au travers du formalisme de diagramme de classes étendu que nous avons présenté dans le second chapitre (classe entrepôt, environnement, filtres temporel et d'archives). Les magasins de données sont représentés à l'aide des représentations graphiques des faits, des dimensions et des contraintes. La fenêtre de la source globale sert de support au processus de construction graphique et incrémentale du schéma de l'ED. De la même façon, le schéma de l'ED sert de support au processus d'élaboration graphique des magasins de données multidimensionnelles.

2.2.3 Stockage

Dans le cadre du projet REANIMATIC, notre partenariat avec des hôpitaux parisiens nous a amené à travailler avec un SGBD relationnel (ORACLE). Aussi, la source, l'ED et les MD sont représentés au travers de tables. Nous avons donc défini des processus de traduction entre les représentations conceptuelles et le stockage des données reposant sur le modèle relationnel.

2.2.4 Traitements

Ce module contient deux générateurs permettant de construire et d'implanter automatiquement un ED ou un MD.

Le générateur d'ED traduit les actions graphiques qu'effectue un concepteur sur une source globale afin de construire automatiquement un ED. La construction d'un ED repose sur trois scripts. Ces scripts correspondent à la création de l'ED dans un SGBD hôte ainsi qu'à l'initialisation (première extraction pour peupler l'entrepôt) et aux rafraîchissements (extractions suivantes pour répercuter les évolutions des données source) de l'entrepôt. Ces scripts reposent sur un processus de traduction. Nous avons proposé un ensemble de règles de passage du niveau conceptuel objet vers le niveau logique relationnel. Nous retrouvons des règles permettant de transformer les concepts objet standards (classe, héritage, association, agrégation et attributs multivalués), et des règles permettant de transformer les concepts inhérents à notre modèle d'entrepôt (états passés et archivés). Les classes d'un environnement sont traduites par 3 relations : relations des états courants, passés et archivés.

Le générateur de MD traduit les actions graphiques qu'un concepteur effectue sur un schéma d'ED pour la génération et l'implantation d'un MD dans un SGBD relationnel. Pour répondre à ce besoin, nous avons utilisé les règles de transformation d'un schéma multidimensionnel en schéma R-OLAP. Toute dimension est traduite en une table dont les attributs correspondent aux paramètres et attributs faibles. La clé primaire d'une table de dimension correspond au paramètre de plus faible granularité. Tout fait est traduit par une table dont les attributs rassemblent les mesures et les clés étrangères reliées aux clés primaires des dimensions liées au fait.

2.3 ELABORATION GRAPHIQUE ET INCREMENTALE D'UN ENTREPOT

La construction graphique et l'implantation d'un ED sont réalisées à l'aide du module GEDOOH (Générateur d'Entrepôts de Données Orientés Objet et Historisés). Après avoir choisi le graphe d'une source globale, la démarche proposée par GEDOOH suit trois étapes :

1. extraction des données source utiles pour les décideurs et organisation de ces données dans l'entrepôt au travers des fonctions de construction (définissant les classes de l'entrepôt),

2. historisation des données en définissant des environnements ainsi que les filtres associés aux classes,

3. configuration des données par l'intermédiaire de règles qui déterminent le comportement de l'entrepôt (période de rafraîchissement, critères d'archivage…).

Chacune de ces étapes est présentée de manière détaillée dans [Teste, 2000]. Dans les sections suivantes, nous mettons en avant les particularités de celles-ci.

2.3.1 Extraction et organisation

Les **fonctions d'extraction** sont directement exprimées sur le graphe de la source. L'administrateur exprime une requête graphique (sélection de nœuds ou liens sources et utilisation des menus). Toute fonction d'extraction exprimée graphiquement correspond à une suite d'opérations algébriques exécutables (processus de "mapping" défini dans le second chapitre). Le résultat d'une opération est une (ou plusieurs) classe(s) intermédiaire(s) représentée(s) par un ensemble de nœuds et de liens affichés en inverse vidéo pour permettre à l'administrateur de mesurer la portée de son opération, mais également pour lui permettre de poursuivre par une opération suivante (processus d'élaboration incrémentale).

Les **fonctions d'organisation** permettent d'adapter le graphe de l'entrepôt aux besoins de l'administrateur. Les opérations de réorganisation permettent de supprimer des classes, des liens, des attributs, des opérations mais également de créer des super-classes et des sous-classes.

Exemple. Dans la figure suivante, vous trouvez les représentations graphiques de la source et de l'entrepôts ainsi qu'un exemple de fonction d'extraction suite à la selection dans l'entrepôt de la classe Etablissement.

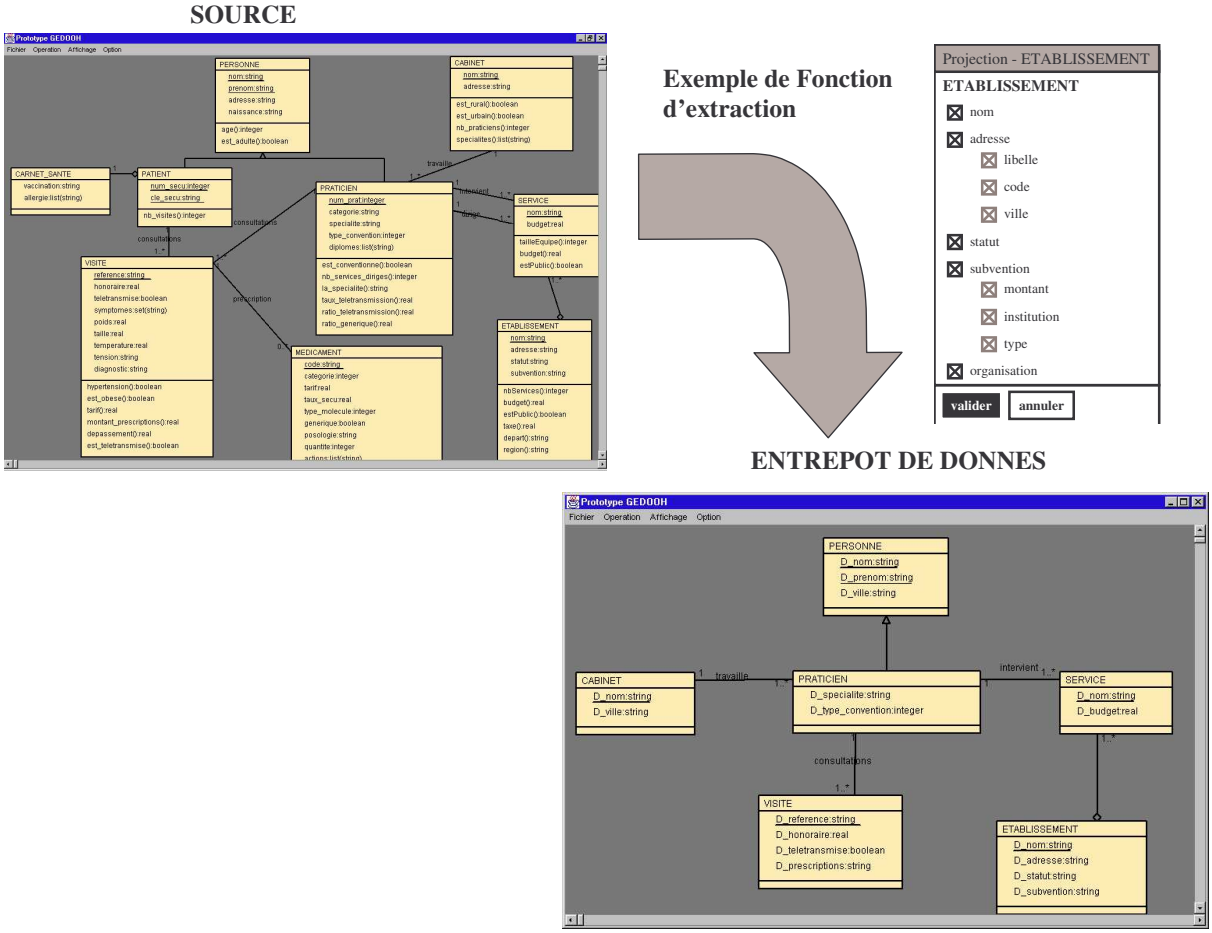


Figure 63 : Exemple de fonction d'extraction

2.3.2 Historisation

La seconde étape de notre démarche consiste à se focaliser sur les aspects temporels de l'entrepôt. L'administrateur définit les différents environnements et les filtres associés.

Pour définir graphiquement une partie historisée, l'administrateur sélectionne directement sur le graphe de l'entrepôt les classes qu'il souhaite placer dans un environnement. Après avoir déclenché l'opération de création d'un environnement, il spécifie ensuite le nom de celui-ci et par l'intermédiaire de menus les filtres temporels et d'archives

Exemple. Après avoir construit un environnement contenant les classes "service" et "établissement", l'administrateur définit les attributs pour lesquels, il souhaite conserver les états passés (filtre temporel).

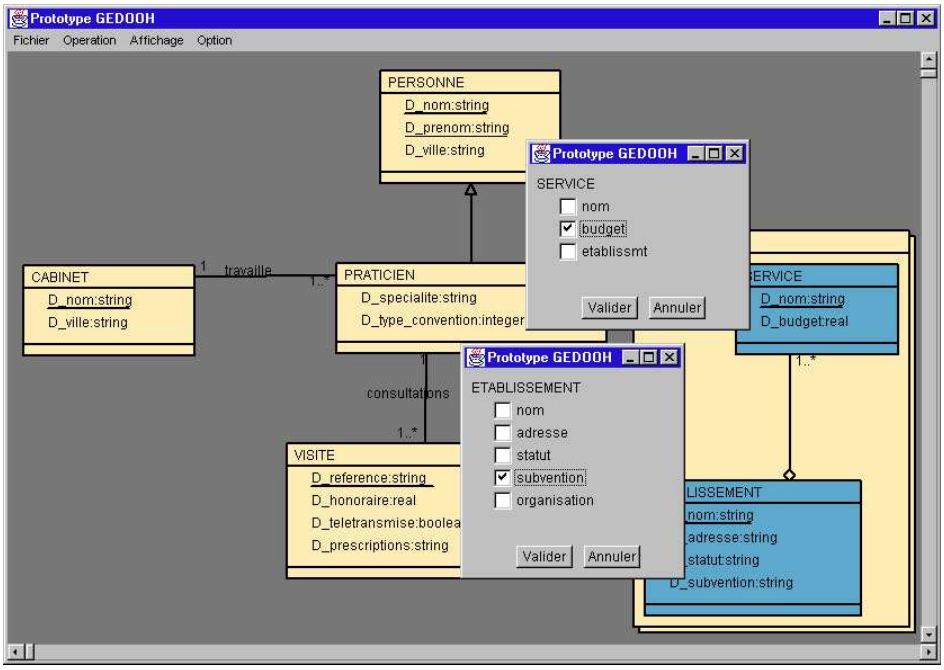


Figure 64 : Exemple de définition d'un filtre temporel

2.3.3 Configuration

Cette dernière étape permet de caractériser le comportement de l'entrepôt et des environnements en définissant des règles fixant des périodes de rafraîchissement, des critères d'archivage des différents environnements (déterminant les états passés à archiver)… Se reporter au chapitre II pour plus de détails.

2.4 ELABORATION DE MAGASINS MULTIDIMENSIONNELS CONTRAINTS

Ce module, appelé GMAG (Générateur de MAGasins de données) assiste le concepteur dans la définition du schéma multidimensionnel du magasin à partir d'un entrepôt de données historisées. Le schéma multidimensionnel obtenu est basé sur notre modèle en constellation à base de contraintes sémantiques. Ce module permet de valider la démarche ascendante que nous avons définie dans le chapitre précédent.

L'élaboration des magasins de données repose donc sur les 5 étapes suivantes :

- détermination des faits représentant les sujets analysés,
- détermination des dimensions représentant les perspectives de l'analyse,
- définition de la granularité des données de l'analyse,

- organisation des paramètres des dimensions selon des dépendances hiérarchiques pour supporter les analyses à différents niveaux de détail,
- expression des contraintes sémantiques inter et intra-dimension.

Les différentes étapes de cette démarche ont été explicitées dans le chapitre précédent. L'implantation de celle-ci et son utilisation au travers de l'outil G-MAG sont détaillées dans [Ghozzi, 2004]. Les paragraphes suivants visent à présenter quelques exemples d'interfaces de ce logiciel.

Exemple. : Dans la figure ci-dessous, le schéma d'un ED est représenté en arrière plan. La classe entrepôt "Location " a été sélectionnée afin de servir de support à la construction d'un fait intitulé "Loc_Vehicule". Pour l'élaboration de ce dernier, l'administrateur choisit les mesures désirées et les fonctions d'agrégations associées.

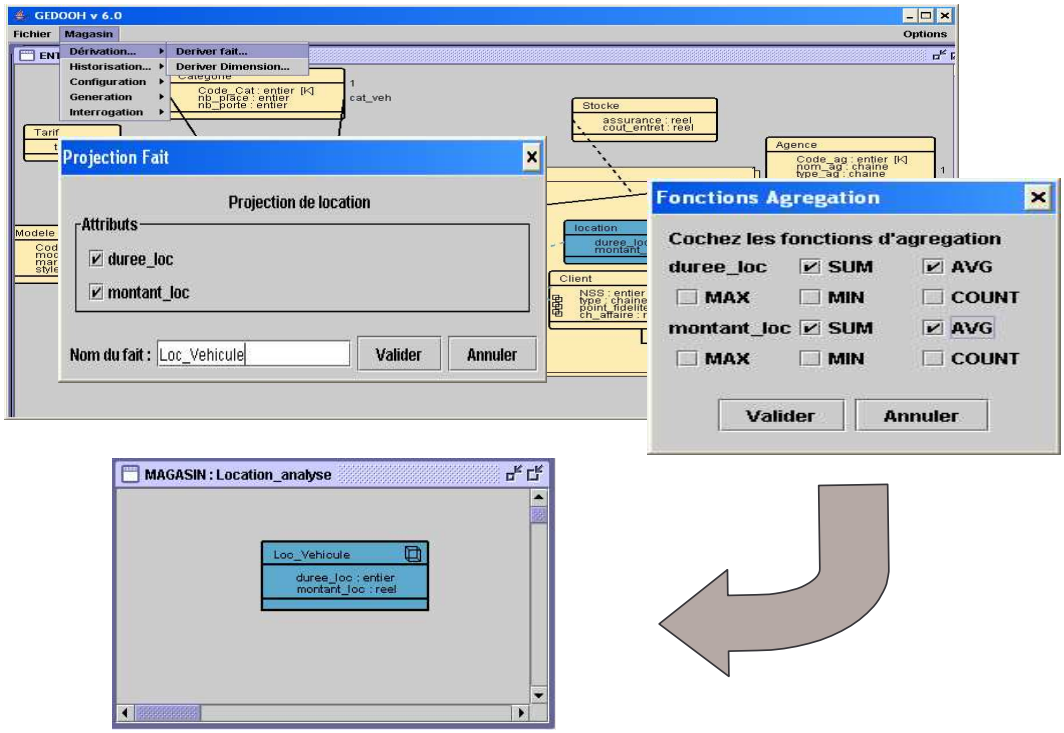


Figure 65: Dérivation du fait « Loc_vehicule »

Exemple. La construction des dimensions repose sur la sélection d'attributs de classes déterminantes. Cette sélection s'effectue au travers de menus comme pour les faits. La définition complète des dimensions nécessite la description de sa hiérarchisation. La figure suivante donne un exemple de la hiérarchisation de la dimension "Agence" associé au fait défini dans l'exemple précédent. Dans cette figure, nous montrons les étapes à suivre pour construire une hiérarchie et l'affichage final pour la dimension "Agence" contenant trois hiérarchies.

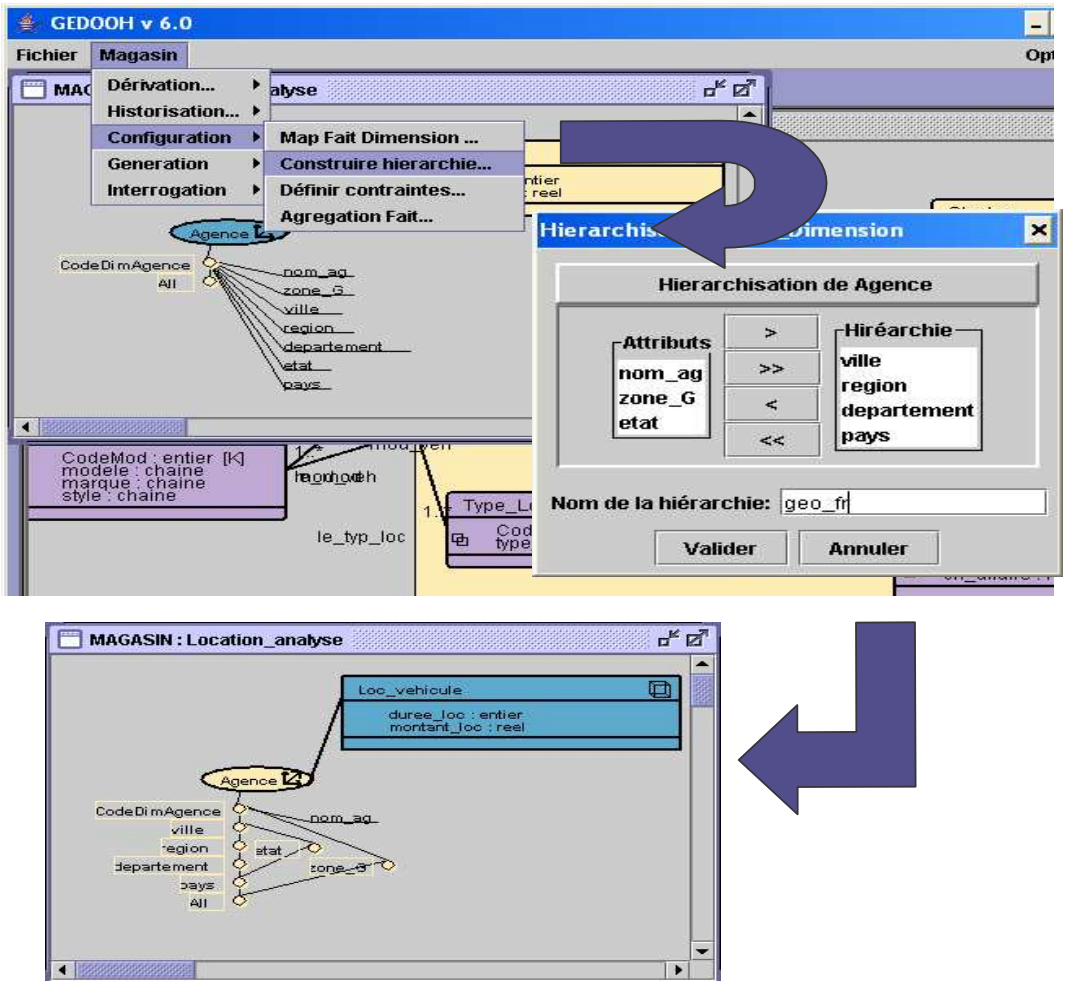


Figure 66 : Hiérarchisation de la dimension "Agences"

3 OUTIL DE MANIPULATIONS MULTIDIMENSIONNELLES

Cet outil permet de manipuler une base de données multidimensionnelles, soit à l'aide de manipulations graphiques (GOLAP), soit à l'aide de notre langage OLAP-SQL. La restitution des données s'effectue sous la forme d'une table multidimensionnelle.

Tous les détails d'implantation et d'utilisation de ces deux langages sont développés dans [Tournier, 2004] et [Annoni, 2003].

Dans la section suivante, nous présentons l'architecture générale de cet outil. Dans les deux autres sections, nous étudions le fonctionnement des langages de manipulations textuelles et graphiques.

3.1 ARCHITECTURE

Cet outil comprend trois niveaux que nous pouvons schématiser comme suit :

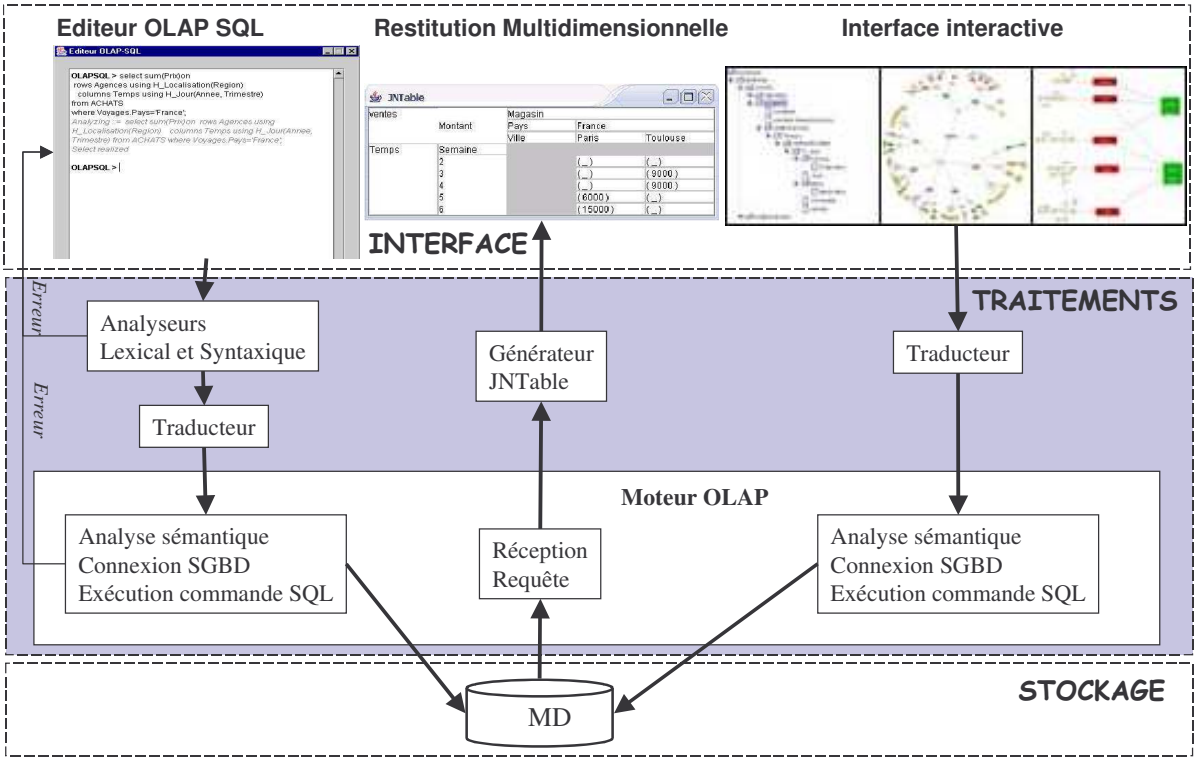


Figure 67 : Architecture de notre outil de manipulation multidimensionnelle

Le niveau **"Interface"** regroupe trois composants. Le premier, l'éditeur de commandes permet de saisir textuellement une commande du langage assertionnel OLAP-SQL et éventuellement d'afficher un message d'avertissement en cas d'une erreur de saisie. Le résultat d'une commande de consultation se traduit par l'affichage des résultats dans une table multidimensionnelle. Le troisième composant permet d'afficher le schéma d'une BD multidimensionnelles sous la forme d'un graphe interactif sur lequel le décideur effectue ses analyses décisionnelles. Ces manipulations graphiques correspondent aux commandes du langage GOLAP.

Le niveau **"Traitements"** comprend différents composants. Toute commande OLAP-SQL est (1) vérifiée lexicalement et sémantiquement, (2) traduite en une commande SQL (3) prise en charge par le moteur OLAP qui l'analyse sémantiquement (consultation de la métabase) et se connecte au SGBD afin de pouvoir exécuter la requête. De même, toute commande graphique GOLAP est (1) traduite en une requête SQL et (2) prise en charge par le moteur OLAP.

Le niveau stockage regroupe deux bases de données. La métabase regroupe toutes les informations (métadonnées) sur les différents composants du magasin de données analysé. Le magasin de données est une base de données R-OLAP stockant le contenu des faits et des dimensions.

3.2 LANGAGE ASSERTIONNEL

L'utilisateur saisit une commande dans l'éditeur de commandes OLAP-SQL. En cas d'erreur lexicale, syntaxique ou sémantique, le système renvoie un message d'erreur. Si le décideur saisit une requête de consultation, le logiciel retourne la table multidimensionnelle correspondante. La figure ci-dessous illustre ce principe. Vous trouverez tous les détails de cet outil dans [Annoni, 2003].

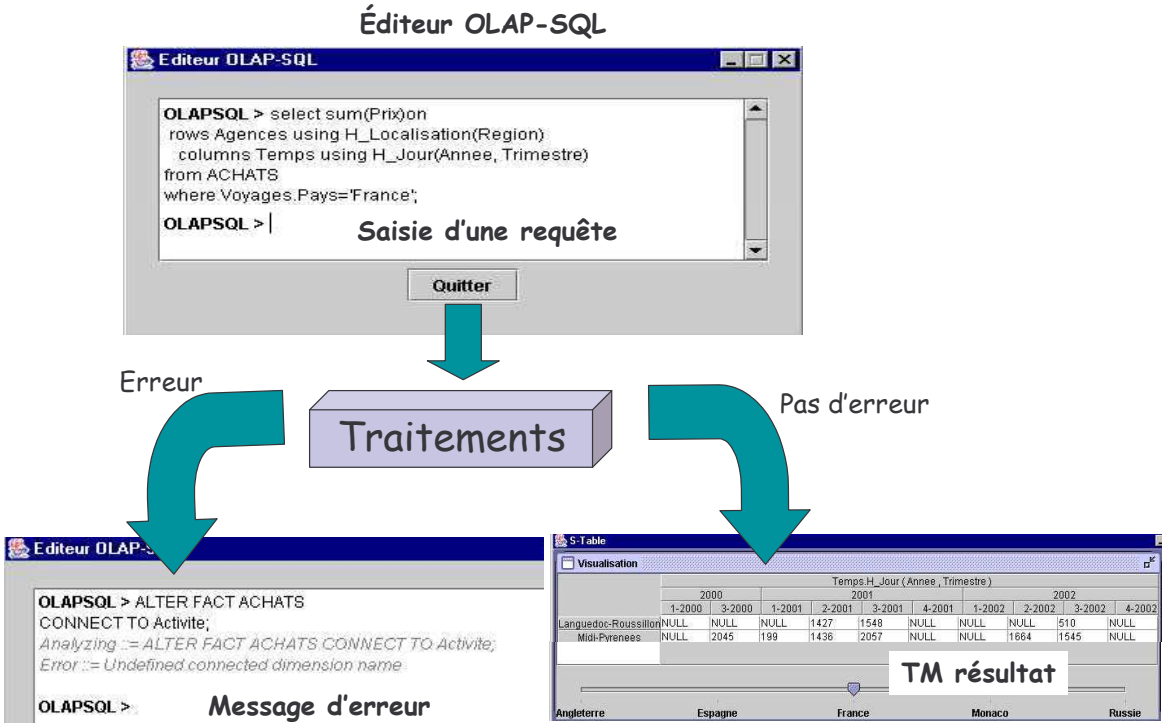


Figure 68 : Exemple de commande assertionnelle

3.3 LANGAGE GRAPHIQUE

3.3.1 Interface de visualisation

Cette interface de visualisation permet de représenter le schéma conceptuel de la BD multidimensionnelles étudiée. Afin d'offrir la meilleure vision possible au décideur, nous avons décliné cette interface graphique en trois versions : arborescente, graphique et hyperbolique.

La **visualisation arborescente** correspond à la représentation la plus usitée par les utilisateurs d'outils informatiques. Elle n'est pas nécessairement la plus adaptée pour des manipulations multidimensionnelles et génère des redondances d'informations. Un fait regroupe ses mesures et ses dimensions liées. De même, une dimension rassemble ses hiérarchies, ses paramètres, ses attributs faibles et ses faits liés.

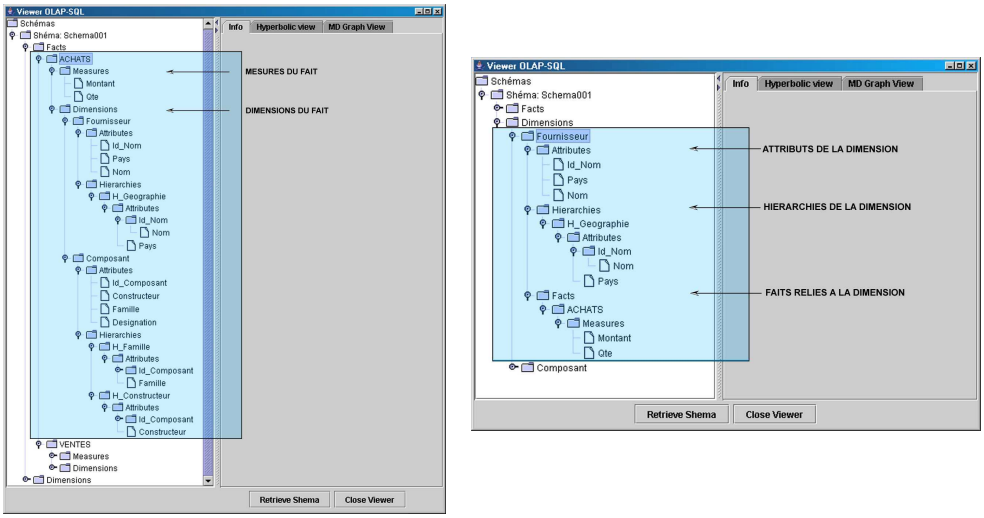


Figure 69 : Exemple de visualisation en arborescence

La **visualisation graphique** repose sur les formalismes du modèle que nous avons défini dans le chapitre 3. Elle repose sur un formalisme permettant d'identifier clairement les différents composants d'un schéma multidimensionnel. Le fait et ses mesures sont représentés par un rectangle vert. Les dimensions sont représentées à l'aide d'un rectangle rouge. Les hiérarchies sont modélisées au travers d'un graphe dont les points jaunes correspondent à des paramètres. Cette visualisation est la plus utilisée.

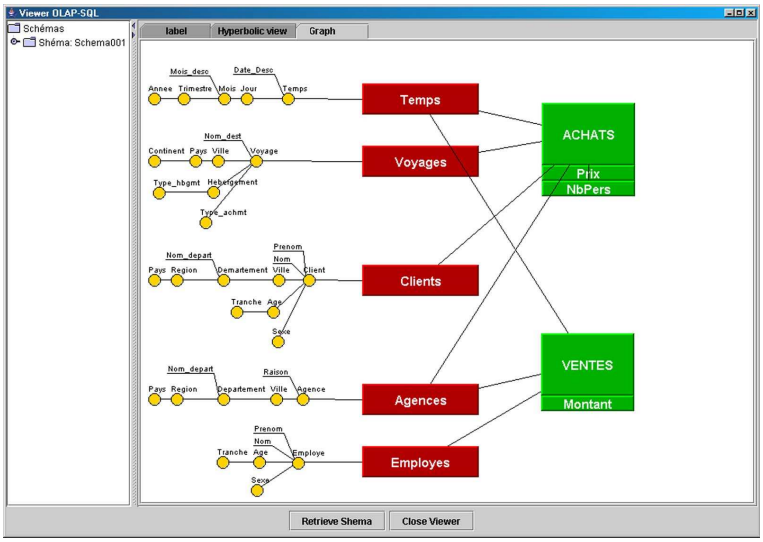


Figure 70 : Exemple de représentation graphique

La **visualisation hyperbolique** ("Fisheye View") est une extension de la représentation graphique présentée dans la section précédente. Elle permet de visualiser et de naviguer de manière plus aisée dans une constellation contenant de nombreuses données. Cette visualisation représente une constellation au travers d'un graphe mettant en valeur la partie étudiée par le décideur [Lamping & Rao, 1994 ; Munzner, 2000]. La figure suivante donne un exemple de cette représentation.

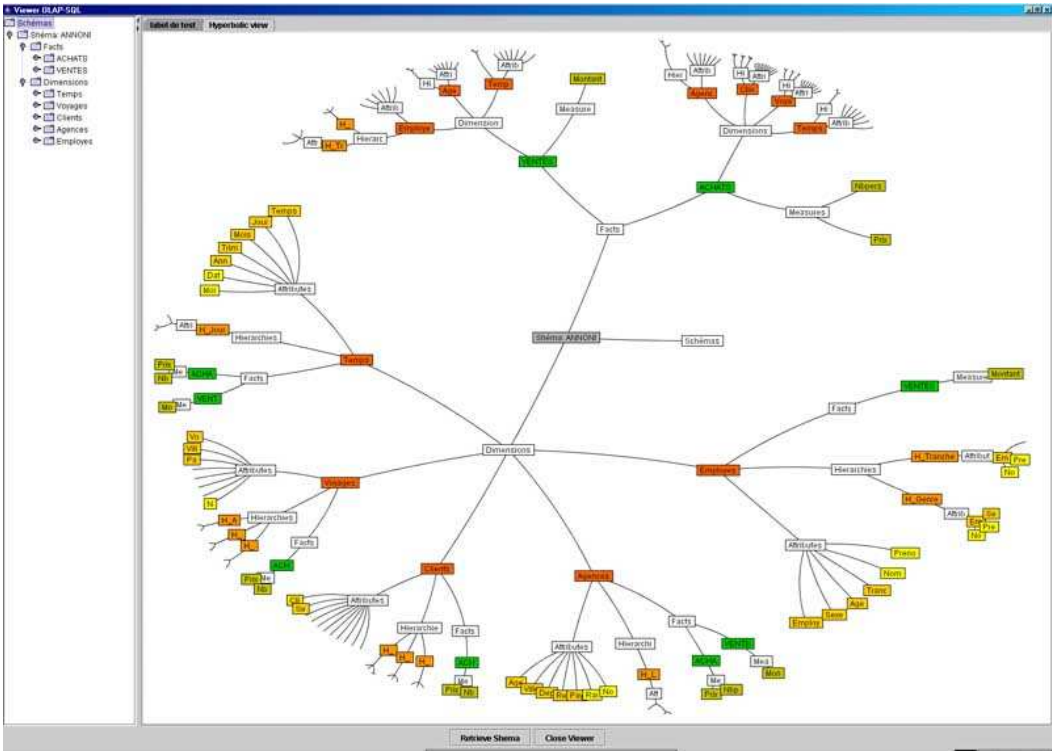


Figure 71 : Exemple de représentation hyperbolique

3.3.2 Interrogation graphique

L'interface de visualisation présentée dans la section précédente sert de support à l'élaboration d'analyses multidimensionnelles. Le décideur peut créer et manipuler des tables multidimensionnelles. Vous trouverez tous les détails d'implantation et d'utilisation de cet outil dans [Tournier, 2003].

Pour afficher une première table multidimensionnelle, le décideur sélectionne un fait et une ou deux dimensions pour faire apparaître des menus contextuels afin de préciser les mesures du fait ainsi que les hiérarchies et éventuellement les paramètres des dimensions à afficher.

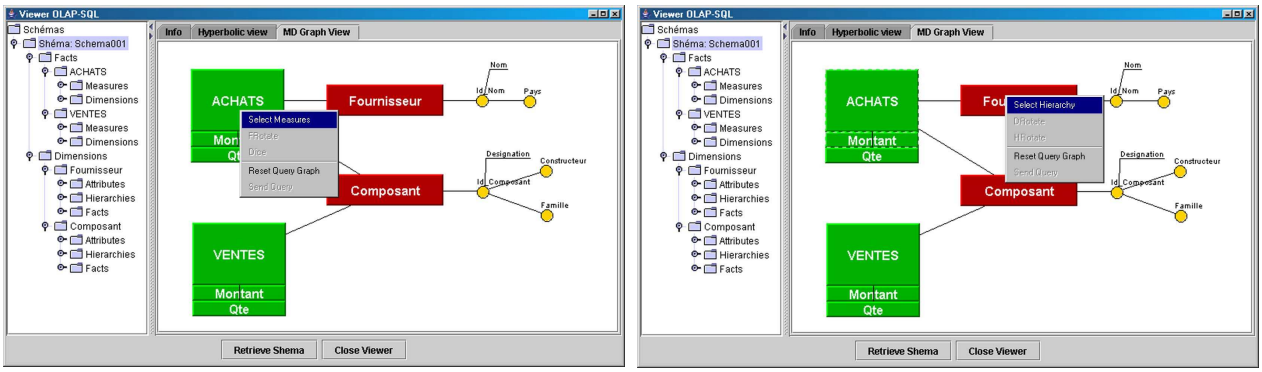


Figure 72 : Exemple de construction d'une TM

Les opérations de l'algèbre multidimensionnelle s'effectuent de manière graphique en cliquant sur un des composants d'un schéma et en sélectionnant une des options contenues dans le menu contextuel affiché.

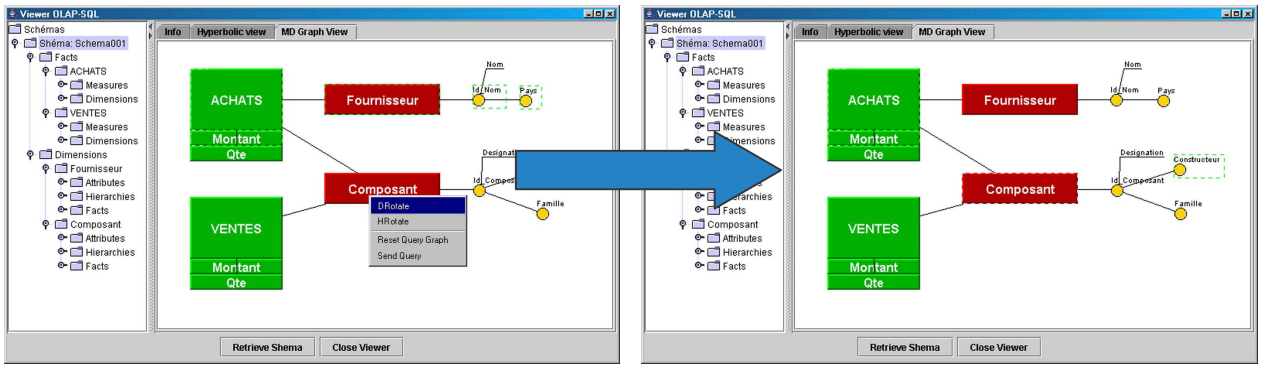


Figure 73 : Exemple de rotation de dimension

4 CONCEPTION ET MANIPULATION D'ENTREPOTS DE DOCUMENTS

Afin de valider les propositions relatives aux entrepôts documentaires, nous avons réalisé un outil d'aide à l'intégration et à l'analyse de documents textuels, intitulé DOCument WAREhouse (DOCWARE). DOCWARE assure les fonctionnalités suivantes :

- supporter une construction graphique incrémentale des entrepôts de documents à partir de documents filtrés et sélectionnés,
- assister le concepteur dans l'élaboration des magasins de documents.

Toutes les fonctionnalités de cet outil sont exposées dans [Khrouf, 2004].

4.1 ARCHITECTURE DE DOCWARE

Cet outil repose sur une architecture à 3 niveaux :

- L'interface graphique permet d'intégrer un document dans un entrepôt documentaire, de visualiser les structures logiques de cet entrepôt et de les manipuler afin de créer des schémas multidimensionnels.
- Le niveau "traitements" regroupe deux modules. Le parseur permet l'intégration automatique des documents XML dans l'entrepôt. Il repose sur les trois phases définies dans le second chapitre : (1) extraction de structure et de contenu, (2) comparaison de structures arborescentes et (3) insertion du contenu du document. Le module analyse permet de construire un schéma multidimensionnel à partir des structures logiques contenues dans l'entrepôt.
- Le niveau stockage permet de sauvegarder les documents dans un SGBD Oracle.

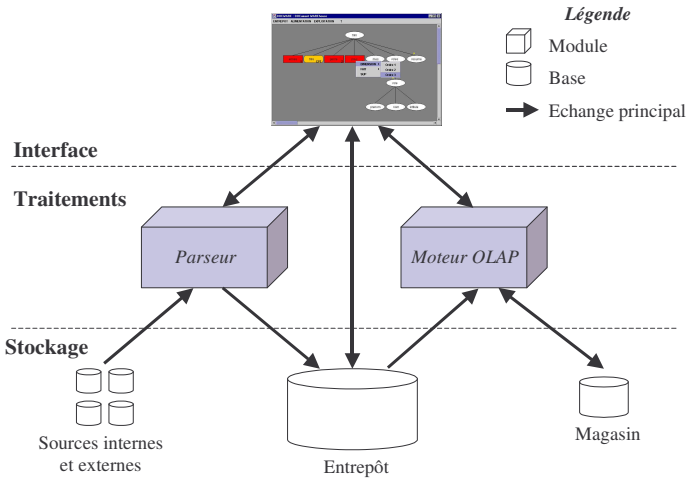


Figure 74 Architecture de DOCWARE

4.2 PARSEUR

L'expérimentation du parseur repose sur l'intégration de **1675 documents**. Ces documents ont été extraits de sources diverses : documents XML issus de sites Web et de CD-ROM fournis dans le cadre de benchmarks (Reuters, …) ou de bases de tests (TREC, …). Ces documents ne sont pas associés à un domaine particulier.

Cette implantation nous a permis de déterminer les valeurs de seuils utiles à notre processus de comparaison. Pour le filtrage consistant à sélectionner les structures logiques génériques semblables à celle du document à intégrer, les expérimentations nous ont permis de déterminer un seuil de filtrage s_f de 0.60. La fusion d'une structure logique spécifique avec la structure logique générique étudiée repose sur la valeur d'un seuil de similarité. Les différents tests que nous avons effectués ont abouti à un seuil de similarité de 0.57.

Dans le cadre de ces expérimentations, nous avons constaté que les classes de documents obtenues étaient homogènes d'un point de vue structurel. Six structures logiques génériques regroupent la majorité des 1675 documents et 17 structures intègrent moins de trois documents. Les six structures logiques génériques les plus significatives ont mis en évidence six classes de documents malgré l'hétérogénéité structurelle initiale des documents à intégrer (news, publications scientifiques, revues, films et tableaux de données numériques). Les 17 structures comportant peu de documents s'expliquent par l'hétérogénéité des sources utilisées et des thématiques associées. Les principales caractéristiques de notre entrepôt sont listées dans le tableau suivant :

Nombre de documents	1675
Nombre de structures logiques génériques	23
Nombre d'éléments génériques	292
Nombre d'attributs génériques	52
Nombre d'éléments spécifiques	61299
Nombre d'attributs spécifiques	19000
Nombre moyen d'éléments par structure	13
Nombre moyen de niveaux de profondeur	4
Nombre moyen de fils par élément père	3
Nombre maximum de fils par élément père ; $\forall \upsilon \in V$, max(Card(fils[υ]))	8
Nombre maximum de documents par structure logique générique	465
Nombre minimum de documents par structure logique générique	1

Tableau 1 : Caractéristiques des expérimentations du parseur

4.3 MOTEUR OLAP

Ce module permet de construire un schéma multidimensionnel et affiche la table multidimensionnelle résultat. Pour la construction du schéma multidimensionnel,

- le système affiche les structures logiques génériques, les éléments génériques, les attributs génériques ou la liste des documents de l'entrepôt
- le concepteur sélectionne un des composants affichés et l'affecte à une structure multidimensionnelle (fait, dimension) via des menus contextuels.

Exemple. Dans la figure suivante, le décideur souhaite analyser le nombre annuel de films sortis par pays et par genre.

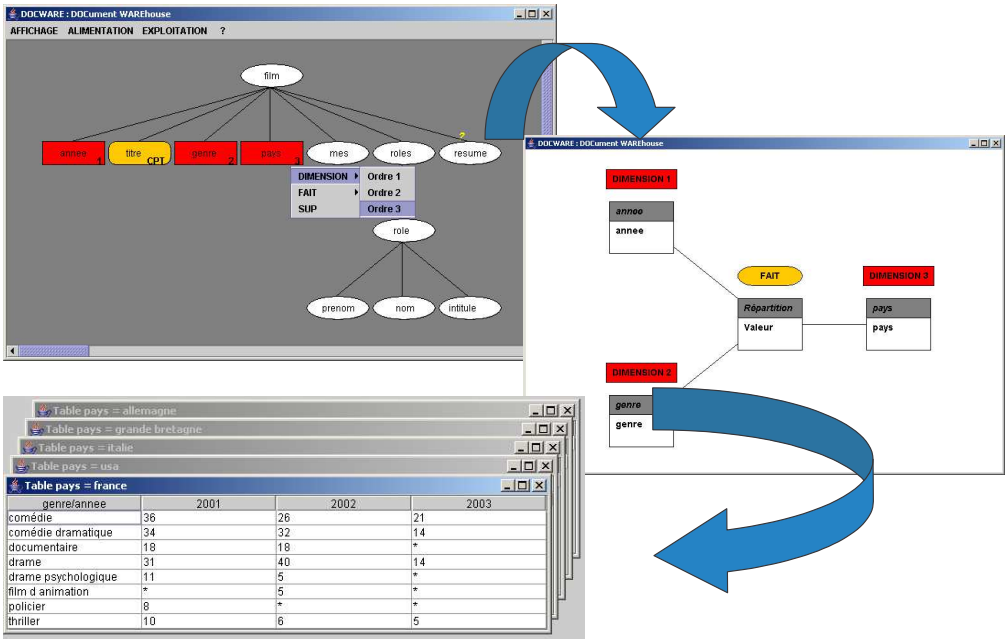


Figure 75 : Analyse multidimensionnelle à partir d'un entrepôt documentaire

5 PROJETS ET COLLABORATIONS

Comme indiqué dans les précédents chapitres, nos travaux ont donné lieu à différentes collaborations avec des entreprises et des laboratoires de recherches. Ces relations se sont traduites par des projets, des conventions de partenariat entre laboratoires de recherches, une bourse CIFRE et des actions incitatives du MENRT.

5.1 VISION SYNTHETIQUE

Afin de mettre en avant les caractéristiques des projets que nous avons menés durant ces dernières années, nous vous proposons le tableau synthétique suivant :

Identifiant	Période	Type	Partenaires	Thème	Rôle personnel
IAPA	2007 - 2011	Projet Laboratoire IRIT	ICR, IBM, CS-SI	Synthèse et analyse croisée des données patient afin de faciliter la découverte de nouveaux diagnostics et de nouvelles thérapies	Co-responsable du lot 6 "Analyses multidimension nelles et décision"
Convention GRIMAAG	2007 - …	Convention de partenariat	GRIMAAG de l'UAG	Processus ETL pour Base de données multidimensionnelles	Responsable scientifique
CIFRE 766/2003	2003-2007	Contrat Bourse CIFRE	I-D6	Méthode de conception de systèmes décisionnels	Co-encadrement d'une thèse
GafoDonnées	2001-2002	Action Spécifique N°20 CNRS-STIC « Fouille de données »	LRI, LIM, LIMOS LORIA, ERIC, LSITT, LIP6, GREYC	Rassembler les communautés des bases de données, de l'apprentissage automatique, de l'analyse de données, et des interfaces hommes machines	Participant
Réanimatic	1999-2000	Action incitative du MENRT	PRISM, IRIT, LAMIH, LIRMM, LIPN, LISI et Outcome REA	Définition d'un outil de conception d'entrepôt de données	Coordinateur IRIT
Evolution	1998-2000	Action incitative du MENRT	PPRISM et IRIT-SIG, avec 13 équipes de recherches	Conception et développement d'un entrepôt de données pour des patients en réanimation	Coordinateur IRIT
Convention CTI-SUD	1997-1998	Convention de recherche	CTI-SUD	Entrepôt de données médicales	Co-responsable

5.2 PRESENTATION DES PROJETS ET/OU COLLABORATIONS

Nous explicitons ces différents projets dans les paragraphes suivants.

IAPA : le projet IAPA (Infrastructure d'Accès, de Partage et d'Analyse de données biomédicales), est un projet des laboratoires IRIT et ICR (Institut Claude Regaud) en partenariat avec deux entreprises (IBM et CS-SI). Ce projet, porté par le professeur J.M. Pierson, se place dans le cadre du pôle de compétitivité "Cancer-Bio-Santé" en réponse à la création du cancéropole de Toulouse. De nos jours, les données médicales sont nombreuses, hétérogènes, temporelles et pas nécessairement structurées. Ce projet vise à concevoir et à développer une infrastructure d'accès, de partage et d'analyses croisées de ces données bio-médicales pour l'aide au diagnostic. Ce projet a été décomposé en 8 lots. Il repose sur des principes d'extraction de

données, de stockage et manipulation de données au sein de grilles de calcul, de modélisation et de recherche d'information, d'extraction de caractéristiques dans des images, d'analyses multidimensionnelles, de mise à disposition des résultats et de routine clinique. Pour l'instant, ce projet a récu un financement partiel (BQR de l'Université Paul Sabatier et fondation InaBioSanté). Nous recherchons encore de nouveaux partenaires financiers (région, fondation InabioSanté…). Pour le lot 6, mon rôle a consisté à définir les axes de recherche (modèles de représentation multidimensionnelle de données hétérogènes, analyses décisionnelles de données hétérogènes, fouille de données hétérogènes), les éléments financiers et les relations avec les partenaires. De plus, une thèse relative à ce lot a débutée en septembre dernier (H. Jerbi).

Convention GRIMAAG : cette année, nous avons également signé une convention de partenariat avec le laboratoire GRIMAAG (Groupe de Recherche en Informatique et Mathématiques Appliquées des Antilles-Guyane) de l'UAG (Université Antilles-Guyanne). Cette convention vise à proposer des solutions dans le développement de Bases de Données Multidimensionnelles (BDM). Notamment, nous souhaitons proposer de nouvelles solutions pour l'intégration des processus ETL lors des phases d'analyse, de conception et de développement de BDM. Ma contribution a consisté à définir les orientations de recherche pour cette collaboration.

CIFRE 766/2003 : la société I-D6[19] a fait appel à notre équipe pour lui apporter des solutions pour l'analyse et la conception de Système d'Aide à la Décision (SAD). Afin de faciliter le turnover inhérent aux SSII, I-D6 souhaite disposer d'un processus fiable et rapide de développement de SAD tout en capitalisant l'expérience et la connaissance de ses collaborateurs. Cette collaboration s'est déroulée sur la base d'une convention de bourse CIFRE pour la thèse d'Estella Annoni (CIFRE 766/2003). Cette thèse [Annoni, 2007] a proposé (1) des modèles et des règles de transformation pour les phases d'analyse et de conception (2) un algorithme pour la définition des modules d'un SAD (3) un catalogue de patrons processus et produits et (4) un outil pour la gestion et la réutilisation des patrons. Dans ce cadre, mon rôle s'est traduit par le co-encadrement de cette étudiante (80%).

GafoDonnées : Cette Action Spécifique (CNRS-STIC AS N°20 : « Extraction et Fouille ») se fondait sur les faits suivants :

- la fouille de données et l'extraction de connaissances constituent un domaine pluridisciplinaire au confluent des bases de données, de l'apprentissage automatique, de l'analyse de données et des statistiques, et des interfaces hommes machines,
- le cloisonnement des recherches est dommageable pour la présence française dans la recherche européenne et internationale.

GafoDonnées a permis de rassembler les communautés françaises en base de données et apprentissage automatique, de dégager des projets de recherche communs et d'approfondir quelques sujets considérés comme prioritaires. Mon implication a notamment consisté en la participation aux groupes de travail GafOLAP et en la co-rédaction du rapport final de ce groupe : "Entrepôts de données et OLAP : un aperçu orienté recherche" [Laurent et al., 2002].

REANIMATIC : ce projet national a été financé par le Ministère de l'Education Nationale, de la Recherche et de la Technologie à hauteur de 1 MF. Le projet REANIMATIC s'effectua en collaboration avec l'association de médecins OUTCOME-REA[20] et plusieurs équipes de recherche (PRISM, IRIT, LAMIH, LIRMM, LIPN, LISI). Jusque là, aucun outil ne permettait d'aider le clinicien à prédire la survenue de complications nosocomiales des patients en réanimation. Pourtant de nombreuses données sont collectées chaque jour auprès des malades.

[19] SSII spécialisé dans le domaine des systèmes d'aide à la décision (http://www.i-d6.com/)
[20] http://www.outcomerea.net/

Le but de ce projet est d'élaborer un outil d'aide à la décision fondé sur un entrepôt de données. Notamment, ce projet a proposé une modélisation spécifique des entrepôts de données médicales évolutives (collectées à partir des bases opérationnelles des services de réanimation) afin d'améliorer la qualité des soins et le devenir des patients dans les services de réanimation des hôpitaux français. Mon action a consisté à participer et à organiser des réunions de travail ainsi qu'à définir et à organiser les différentes tâches à réaliser pour l'équipe SID-ED de l'IRIT. Ce projet a servi notamment de cadre applicatif aux travaux de thèse d'Olivier Teste [Teste, 2000] que j'ai encadré à 80%.

EVOLUTION : le projet EVOLUTION, soumis dans le cadre des Actions Incitatives en 1997 par l'équipe SIG de l'IRIT et le laboratoire PRISM de l'Université de Versailles-Saint-Quentin, a fédéré des équipes de recherche françaises dans le domaine des entrepôts de données. Le Ministère de l'Education Nationale, de la Recherche et de la Technologie a financé ce projet à hauteur de 100KF. Ce projet a permis de poser les bases pour le développement de systèmes d'aide à la décision (architecture modulaire, modélisation conceptuelle, logique et physique d'ED et de MD) ainsi qu'à son exploitation. Le résultat de cette action s'est traduit par l'écriture d'un chapitre de livre par le groupe Evolution [Evolution, 2001]. Comme pour le précédent projet, mon action s'est matérialisée par la participation et l'organisation de réunions ainsi que la définition des tâches à réaliser par notre équipe.

Convention CTI-SUD : cette collaboration s'est déroulée dans le cadre d'une convention de recherche entre l'Université Paul Sabatier (équipe SIG) et le CTI-Sud (Centre de Traitement Informatique de l'Assurance Maladie à Toulouse). Cette étude a eu pour but de spécifier un entrepôt de données médicales alimenté par la base SIAM (Système d'Information de l'Assurance Maladie). Cette collaboration a servi de support à la thèse d'Olivier Teste [Teste, 2000]. Ma contribution s'est traduite par l'encadrement de la thèse d'Olivier Teste à 80%, ainsi que la participation à des activités de gestion de projet (analyse des besoins, suivi des phases d'analyse et de conception, relations avec le partenaire).

6 ENCADREMENTS ET PUBLICATIONS

6.1 ENCADREMENTS

Dans cette première section, vous trouverez une présentation synthétique de mes activités d'encadrement.

Ma contribution s'est traduite par le co-encadrement de 5 thèses dont la dernière sera soutenue en décembre 2007.

Années	Etudiant	Domaine	Type	Rôle personnel
2004-2007	R. Tournier	Intégration de documents XML dans un schéma multidimensionnel	Thèse	Co-Encadrement (40%)
2003-2007	E. Annoni	Méthode d'analyse, de conception et de développement de système décisionnel	Thèse avec Bourse CIFRE n° 766/2003 (Société I-D6)	Co-Encadrement (80%)
2000-2004	F. Ghozzi	Modèle multidimensionnel à contraintes avec démarche et langage de manipulation	Thèse avec allocation MNERT - CIES	Co-Encadrement (80%)
2000-2004	K. Khrouf	Modélisation et manipulation d'entrepôts de document	Thèse	Co-Encadrement (50%)
1997-2000	O. Teste	Modélisation et manipulation d'entrepôts de données historisées	Thèse avec allocation MNERT - CIES	Co-Encadrement (80%)

Depuis l'an dernier, j'ai initié une collaboration avec l'Institut National de Formation en Informatique (INI) d'Alger. Cette collaboration s'est traduite par la proposition d'un cours en ingénierie des système d'information décisionnel ainsi que l'encadrement des travaux de recherches de deux étudiants en Magistère SIC (Système d'Information et de Communication).

Années	Etudiant	Domaine	Rôle personnel
2006-2007	S. Hafyane	Langages graphiques pour bases de données multidimensionnelles temporelles	Encadrement
2006-2007	A. Nassim	Intégration de documents XML dans un schéma multidimensionnel	Encadrement

Depuis 1997, j'ai encadré ou co-encadré 17 étudiants en DEA 2IL (Informatique de l'Image et du Langage) ayant évolué en 2004 sous la forme du Master 2 Recherche 2IH (Image, Information et Hypermédia) de l'Université Paul Sabatier

Années	Etudiant	Domaine	Rôle personnel
2006-2007	H. Jerbi	Mémoire d'expertises décisionnelles à base d'annotations	Co-encadrement
	C. Koussa	Bases de données multidimensionnelles : construction d'un magasin de données à partir de sources XML	Co-encadrement
2005-2006	L. Benakezou	Etude et Conception de bases de données multidimensionnelles temporelles	Co-encadrement
	M. Gargouri	Assistance à l'élaboration incrémentale d'un magasin de données	Co-encadrement
2004-2005	E. Negre	Evolution de schémas dans une constellation	Co-encadrement
	O. Rouhaud	Bases de données décisionnelles : Fusion de tables multidimensionnelles	Co-encadrement
	F. Boucheikh	Méthodologie de conception de systèmes décisionnels	Co-encadrement
	A. Tahi	Interface d'interrogation incrémentale de données multidimensionnelles	Co-encadrement
	M. Kaddes	Etude de faisabilité d'une modélisation en constellation sous contraintes sémantiques	Encadrement
2003-2004	B. K. Le Thi	Intégration de contraintes dans OLAP-SQL	Co-encadrement
	M. Sallami	Développement d'une politique d'accès aux bases en constellation	Co-encadrement
	R. Tournier	Vers un langage de manipulation graphique des bases multidimensionnelles	Co-encadrement
2002-2003	E. Annoni	Conception et développement d'un langage assertionnel pour les bases de données multidimensionnelles	Co-encadrement
2001-2002	L. Bouzguenda	Conception et implantation d'un prototype d'interrogation et de visualisation d'une base multidimensionnelle	Co-encadrement
2000-2001	T. Jarraya	Conception et implantation des mécanismes d'extraction des données sources pour construire un entrepôt de données	Encadrement
1999-2000	F. Ghozzi	Mécanismes d'historisation et d'archivage pour les entrepôts de données complexes	Encadrement
1998-1999	X. Baril	Historisation dans les entrepôts de données	Encadrement

6.2 PUBLICATIONS

Dans cette section, je vous propose la liste des articles classés par année et par type de publication. Pour chacun d'eux, nous avons précisé :

- la référence (pour trouver tous les détails dans la bibliographie),
- le lieu,
- le type (RI pour Revue Internationale, RN pour Revue nationale, CI pour Conférence Internationale, CN pour Conférence Nationale, OI pour Ouvrage International, ON pour Ouvrage National, A pour Autres), et,
- l'objectif de chacun des articles.

2007

Référence	Lieu	Type	Objectif
[Ravat et al., 2007a]	Inter. Journal of Datawarehousing and Mining	RI	*Cet article présente de manière détaillée et complète toutes nos propositions en matière d'algèbre et de langage graphique*
[Cabanac et al., 2007]	DAWAK'07	CI	*Cet article présente un modèle conceptuel de bases de données multidimensionnelles annotées et sa traduction dans un environnement R-OLAP*
[Ravat et al., 2007b]	ICEIS'07	CI	*Cet article présente une fonction d'agrégation pour des données textuelles XML dans le cadre d'analyses multidimensionnelles*
[Ravat et al., 2007c]	SEKE'07	CI	*Cet article présente une démarche pour l'intégration de données issues de documents XML dans un magasin de données multidimensionnel permettant l'analyse de données textuelles*
[Ravat et al., 2007d]	ADBIS'07	CI	*Cet article présente un langage graphique de manipulation de base de données multidimensionnelle reposant sur une algèbre*
[Ravat et al., 2007e]	ER'07	CI	*Cet article présente un modèle multidimensionnel adapté à l'analyse de documents semi-structurés ainsi qu'un ensemble complet d'opérations de manipulation décisionnelle.*
[Annoni et al., 2007]	SEKE'07	CI	*Cet article spécifie l'étude des besoins de la branche tactique du trident décisionnel*
[Ravat et al., 2007h]	Document numérique	RN	*Cet article définit une démarche pour spécifier la hiérarchisation des dimensions d'un schéma multidimensionnel de documents ainsi que son implantation logique*
[Khrouf at al., 2007a]	Revue CID	RN	*Cet article explicite l'intégration de documents multimédias dans un entrepôt au travers d'un modèle générique spécifique*
[Ravat et al., 2007f]	INFORSID'07	CN	*Cet article présente notre modèle de personnalisation de constellation et le langage ECA de définition d'une personnalisation*
[Ravat et al., 2007g]	EDA'07	CN	*Cet article présente une modélisation galaxie de documents pour faciliter les analyses multidimensionnelles*
[Khrouf at al., 2007b]	VSST'07	CN	*Cet article présente un modèle générique à versions pour l'entreposage de versions de documents.*

2006

Référence	Lieu	Type	Objectif
[Ravat & Teste, 2006]	IRECOS	RI	*Cet article présente notre modèle multidimensionnel à versions avec une implantation en R-OLAP*
[Ravat et al., 2006a]	DAWAK'06	CI	*Cet article présente notre modèle conceptuel multidimensionnel à base de versions (versions de fait, de dimension et de constellation) ainsi que les processus d'alimentation entre ces différentes versions*
[Annoni et al., 2006b]	DEXA'06	CI	*Cet article explicite le principe de détermination automatique de l'architecture d'un système décisionnel.*
[Annoni et al., 2006a]	DAWAK'06	CI	*Cet article présente le processus de formalisation des besoins tactiques et stratégiques au travers du concept de diagramme décisionnel ainsi que la fusion de ces différents besoins.*
[Annoni et al., 2006c]	RSTI-ISI	RN	*Cet article propose d'évaluer dès le début de l'analyse le contexte statique (données décisionnelles) et dynamique (traitements ETL et autres) suivant les composantes technique et décision de la démarche du trident décisionnel.*
[Annoni et al., 2006d]	EDA'06	CN	*Cet article présente notre processus de collecte des besoins utilisateurs et de leur formalisation à l'aide de règles de structuration.*
[Annoni et al., 2006e]	INFORSID'06	CN	*Cet article présente l'implantation d'une méthode de conception de système décisionnel au sein de la société I-D6 selon la démarche de la roue de Deming.*
[Ravat et al., 2006b]	INFORSID'06	CN	*Cet article présente notre algèbre de manipulation de données multidimensionnelles et son implantation à l'aide d'un langage graphique*
[Cabanac et al., 2006a]	EGC'06	CN	*Cet article présente l'intégration des annotations dans un modèle multidimensionnel.*
[Cabanac et al., 2006b]	EDA'06	CN	*Cet article présente le méta-modèle que nous avons défini pour les bases de données multidimensionnelles annotées.*
[Chrisment et al., 2006]	Encyclopédie de l'informatique et des systèmes d'information	ON	*Cet article définit les concepts d'entrepôt et de magasin, d'alimentation d'entrepôts à l'aide de médiateurs, de modèles multidimensionnels conceptuel et logique ainsi que d'algèbre multidimensionnelle.*

2005

Référence	Lieu	Type	Objectif
[Ravat et al., 2005a]	Database Modeling for Industrial Data Management	OI	*Cet article présente notre modèle multidimensionnel contraint avec les langages algébriques et assertionnel associés*
[Annoni et al., 2005a]	RSTI-ISI	RN	*Cet article montre l'intérêt et l'utilisabilité des patrons dans notre méthode de conception de système d'aide à la décision*
[Annoni et al., 2005b]	AIM'05	CN	*Cet article définit le concept de trident décisionnel et les différentes étapes le composant.*
[Ghozzi et al., 2005]	EDA'05	CN	*Cet article décrit notre démarche de conception mixte de bases de données multidimensionnelles.*

[Ravat et al., 2005b]	EGC'05	CN	*Cet article se centre sur l'opération de fusion de tables multidimensionnelles*
[Chrisment et al., 2005]	Traités des Techniques de l'Ingénieur	A	*Cet article explicite les concepts d'entrepôts et magasins de données pour être facilement compréhensibles par des ingénieurs en activité.*

2004

Référence	Lieu	Type	Objectif
[Ghozzi et al., 2004]	RSTI-ISI	RN	*Cet article propose une typologie complète des contraintes sémantiques d'un schéma en constellation ainsi que l'algèbre de manipulation de données multidimensionnelles contraintes*

2003

Référence	Lieu	Type	Objectif
[Ghozzi et al., 2003a]	ICEIS'03	CI	*Cet article présente les concepts de notre modèle à contraintes et le prototype les supportant.*
[Khrouf et al., 2003]	RSTI - ISI	RN	*Cet article présente les trois étapes du processus d'alimentation d'un entrepôt documentaire.*
[Ghozzi et al., 2003b]	EGC'03	CN	*Cet article présente notre modèle à contraintes et étude les répercussions de ces contraintes sur les opérations de forage et de rotation.*

2002

Référence	Lieu	Type	Objectif
[Ravat et al., 2002]	RSTI-ISI	RN	*Cet article présente les langages de définition, de manipulation et de contrôle des données multidimensionnel (OLAP-SQL)*
[Laurent et al., 2002]	Rapport final GaFOLAP	A	*Ce rapport définit les concepts de BD dans un système d'aide à la décision et de modèles multidimensionnels ainsi que les travaux de recherches en cours.*

2001

Référence	Lieu	Type	Objectif
[Ravat & Teste, 2001]	BDA'01	CN	*Cet article présente notre algèbre de manipulation d'objets et de classes entrepôt.*
[Ravat et al., 2001]	EGC'01	CN	*Cet article présente notre modèle multidimensionnel de base (fait, dimension, hiérarchies)*

2000

Référence	Lieu	Type	Objectif
[Mothe et al., 2000]	ECIS'00	CI	*Cet article propose un modèle de base de données pour le stockage de données issues du Web*
[Ravat & Teste, 2000a]	Entreprise Information Systems II	OI	*Cet article présente les concepts d'objet entrepôt, classe entrepôt, filtre d'archives, filtre temporel et environnement*
[Ravat & Teste, 2000b]	DEXA'00	CI	*Cet article décrit les concepts d'un modèle d'entrepôt et les fonctions de mapping permettant de l'alimenter*
[Ravat & Teste, 2000c]	ADBIS'00	CI	*Cet article explicite l'intégration de pages Web dans un entrepôt de données historisées*

[Ravat et al., 2000]	BDA'00	CN	*Cet article présente notre modèle d'entrepôt de données avec intégration des processus d'extraction de la structure et du comportement des données sources*

1999

Référence	Lieu	Type	Objectif
[Ravat et al., 1999]	CIKM'99	CI	*Cet article propose une extension d'un modèle objet adapté aux entrepôts de données et un langage de configuration de l'entrepôt*

7 BILAN ET SYNTHESE

Dans ce chapitre, nous avons présenté les différentes productions qui nous ont permis de mettre en œuvre et de valider les concepts et principes proposés dans les chapitres précédents.

Dans un premier temps, nous avons proposé un véritable outil d'aide à la conception d'entrepôt et de magasin de données. Cet outil graphique présente l'avantage d'offrir une représentation conceptuelle adaptée des différents espaces de stockage : source, entrepôt et magasins. Ces représentations graphiques servent de support à l'élaboration incrémentale d'un entrepôt et de ses magasins. Suite à la définition graphique d'un entrepôt, le module "Générateur ED" construit la BD relative à l'entrepôt, l'alimente et si nécessaire le rafraîchit. Pour répondre aux spécificités de nos projets, les entrepôts sont stockés dans un SGBD relationnel et nous avons défini des processus de traduction des objets entrepôt, des filtres et des environnements en concepts relationnels. Pour le module "Générateur MD", il repose également sur des scripts de construction d'un MD R-OLAP, d'alimentation et de rafraîchissement. Ce prototype est écrit en Java et Oracle. Il comporte plus de 10000 lignes de code.

Le second prototype permet de manipuler avec les langages graphique et assertionnel une base de données multidimensionnelles. Ce prototype nous a permis de valider nos propositions relatives aux langages de manipulation de données multidimensionnelles et de vérifier la complétude de ceux-ci. Ce prototype est écrit en Java et Oracle. Il comporte plus de 7000 lignes de code.

Le troisième prototype est centré sur l'intégration et la manipulation de documents dans un entrepôt documentaire. Il nous a permis notamment de valider notre processus d'alimentation d'EDO. Ce prototype est écrit en Java et Oracle et l comporte plus de 3000 lignes de code.

Nos travaux se sont intégrés dans le cadre de projets. Ces derniers nous ont permis de collaborer aussi bien avec d'autres laboratoires de recherche que des entreprises ou des associations. Ces projets ont permis de donner un cadre applicatif à certaines thèses que j'ai encadrées. Comme indiqué en section 5, ces projets m'ont également permis d'assurer différentes fonctions : participant, coordinateur IRIT , responsable scientifique voire co-responsable.

D'un point de vue quantitatif, j'ai encadré ou co-encadré 5 thèses, 2 magistères de l'INI, et 17 masters recherche ou DEA de l'Université Paul Sabatier.

Enfin, ces travaux se sont concrétisés par un ensemble de publications que nous pouvons lister comme suit :

- 2 articles dans des revues internationales avec comité de sélection (IJDWM, IRECOS),
- 2 chapitres dans des ouvrages internationaux avec comité de sélection ("Entreprise Information Systems II" – sélection des meilleurs articles de la conférence et "Database modeling for industrial data management"),

- 6 articles dans des revues nationales avec comité de sélection (5 RSTI-ISI et 1 document numérique),
- 14 articles dans des conférences internationales avec comité de sélection (ER, DAWAK, CIKM, DEXA, ICEIS, SEKE, ADBIS, ECIS),
- 14 articles dans des conférences nationales avec comité de sélection (INFORSID, BDA, EDA, EGC),
- 1 article dans l'encyclopédie de l'informatique et des systèmes d'information,
- 1 traité des techniques de l'ingénieur,
- 1 rapport final de projet.

Parmi les 2 revues et les 2 ouvrages internationaux, nos travaux ont été validés par un article dans la revue internationale de référence du décisionnel (International Journal of Datawarehousing and Mining). Au niveau des conférences internationales avec comité de sélection, nous avons 5 articles dans des conférences internationales spécialisées dans le décisionnel (3 articles dans DAWAK et 2 dans SEKE) sans oublier les conférences internationales de renom telles que ER et CIKM. De plus, notre dernier article de DAWAK a été retenu parmi les 10 premiers pour faire l'objet d'une version étendue dans un ouvrage international.

Notre volonté a également été de faire reconnaître nos travaux au niveau national. Cette volonté s'est traduite par 6 articles dans des revues nationales avec comité de sélection (revues RSTI-ISI). Nos travaux ont également été validés par 14 articles dans des conférences nationales avec comité de lecture (4 articles à EDA, 3 articles à INFORSID, 4 articles à EGC, 2 articles à BDA, 1 article à AIM).

La synthèse de ces différents travaux se trouve dans le tableau suivant.

Thèmes[21]	**Projets / collaborations**[22]	**Publications**[23]	**Co-encadrements**[24]	**Participants**[25]
Modélisation d'entrepôts		1 ON (30%)		2 PR, 1 MC
– Entrepôts de données	Reanimatic, CTI SUD	1 OI, 3CI, 2CN	1 thèse, 3 DEA	
– Entrepôts de documents		1 CI, 1 RN	1 thèse	
Modélisation de magasins de données	GafoDonnées, Evolution	1 ON (70%)		1 PR, 3 MC
– Modèle générique de base		1 CN		
– Intégration de données textuelles		2 CI, 1 RN, 1 CN	1 thèse (50%), 1 M2 R, 1MA	
– Gestion de la cohérence sémantique		1 OI, 1 CI, 1 RN, 1 CN	1 thèse (40%), 1 M2 R	
– Gestion de la cohérence temporelle		1 RI, 1CI	2 M2 R	
– Intégration et capitalisation de l'expertise		1 CI, 2 CN	1 M2 R	
– Personnalisation des magasins de données		1CN	1 M2 R	
Manipulation de données multidimensionnelles	IAPA	1 RI, 1 CN		1 PR, 1MC
– Algèbre multidimensionnelle		1 CI	1 thèse (30%)	
– Langage graphique GOLAP		1 CI, 1 CN	1 thèse (50%), 4 M2 R, 1MA	
– OLAP SQL		1 RN	2 M2 R	
Démarche de conception				1 PR, 2MC
– Conception de schémas multidim. contraints		1 CN	1 thèse (30%), 1 M2 R	
– Conception d'un système d'aide à la décision	I-D6, GRIMAAG	3 CI, 2 RN, 3 CN	1 thèse, 1M2R	
Synthèse	7 projets et/ou collaborations	2 RI, 2 OI, 14 CI, 6 RN, 14 CN, 1 ON	5 thèses, 17 DEA ou M2 R, 2 MA	

Tableau de synthèse des thématiques de recherches, des projets, des publications

[21] Thématiques de recherches abordées dans ce mémoire de HDR

[22] Nom des projets et/ou collaborations dont les détails sont donnés en section 5 de ce chapitre

[23] Type des articles (RI pour Revue Internationale, RN pour Revue nationale, CI pour Conférence Internationale, CN pour Conférence Nationale, OI pour Ouvrage International, ON pour Ouvrage National, A pour Autres) avec la proportion en % relative à la thématique

[24] Type d'encadrement (thèse, DEA pour le DEA 2IL de l'Université Toulouse III, M2 R pour le Master 2 Recherche 2IH de l'Université Toulouse III, MA pour le Magsitère SIC de l'INI d'Alger) avec la proportion en% relative à la thématique

[25] Autres permanents ayant collaboré aux travaux de ces thématiques (PR pour Professeur des Université et MC pour Maître de Conférences)

CONCLUSION ET PERSPECTIVES GENERALES

1 BILAN DE NOS TRAVAUX

Les travaux que nous avons menés durant ces dernières années se situent dans le cadre des Systèmes d'Aide à la Décision (SAD). Un SAD permet d'extraire des données de sources de production, pour les agréger et les transformer afin d'être facilement accessibles par les décideurs.

Dans les années 95, les premières recherches relatives aux SAD ont apporté des propositions d'ordre technique avec notamment le concept de vue matérialisée permettant d'extraire les données sources à l'aide d'une requête et de stocker le résultat dans des tables cibles. Ces travaux de recherche se sont essentiellement concentrés sur les phases d'extraction, transformation et chargement des données (ETL). En parallèle à ces travaux de recherches, de nombreux éditeurs de logiciels ont proposé des outils facilitant la prise de décision. Notamment, nous pouvons citer les outils ETL et les outils de requêtage graphique permettant de faire abstraction des caractéristiques d'implantation des données.

Suite à ces différentes propositions, un des problèmes majeurs auxquels sont confrontés les concepteurs est le manque d'outils méthodologiques spécifiques leur permettant d'analyser les besoins utilisateur, de concevoir les applications décisionnelles et de les implanter. En effet, par opposition aux applications de gestion, les SAD doivent intégrer la liaison avec les données sources hétérogènes et réparties, proposer une gestion des évolutions de valeurs dans le temps, et supporter une modélisation facilitant les interrogations et les analyses décisionnelles. Les méthodes de conception dédiées aux applications de production n'étaient donc pas adaptées pour le développement d'applications décisionnelles [Golfarelli & Rizzi, 1998]. En effet, l'élaboration d'un SAD reste une tâche complexe car elle nécessite l'analyse des besoins de différentes unités organisationnelles d'une entreprise [Bruckner et al., 2001b ; Mazon et al., 2005] et constitue une discipline nouvelle n'offrant pas des stratégies et des techniques reconnues [List et al., 2002].

Nos travaux de recherches visent à répondre à ce manque. Plus précisément, les travaux présentés dans ce mémoire apportent des outils méthodologiques pour le développement de SAD à base d'entrepôts de données (ED). Ce choix s'est imposé par le fait qu'un ED est maintenant considéré comme le composant essentiel [List et al., 2002 ; Shim et al., 2002] qui garantit la meilleure réponse aux problématiques décisionnelles des différents domaines fonctionnels d'une entreprise [Franco & De Lignerolles, 2000].

Dans un premier temps, nous avons identifié quatre problématiques. La première consiste en une préparation des données décisionnelles, la seconde vise à la présentation des données aux décideurs, la troisième se centre sur l'exploitation des données multidimensionnelles au travers d'analyses exploratoires et la dernière vise à apporter une démarche de conception et développement de SAD. Ces différentes problématiques reposent sur les concepts d'ED et MD maintenant unanimement reconnus par la communauté.

Préparation de données décisionnelles

Même si de nombreuses personnes sont d'accord sur le fait que l'ED doit permettre de centraliser et historiser l'ensemble des données décisionnelles, il n'y avait pas jusqu'à ce jour de proposition conceptuelle pour la modélisation d'un tel espace de stockage de données. Aussi, nos premiers travaux de recherche ont permis de répondre à ce besoin. Dans un premier temps, nous avons proposé un modèle objet étendu permettant de gérer l'évolution des données au travers de classes contenant des filtres temporels, des filtres d'archives et des fonctions de mapping pour conserver le lien avec les sources. Cette première proposition se limitait aux données issues de BD éventuellement hétérogènes. Cependant, les ED construits uniquement à partir de BD sources n'extraient que 20% des informations disponibles alors que les 80% restants sont stockés dans des documents [Tseng & Chou, 2006]. Aussi, nous avons voulu proposer une solution pour l'intégration de documents au sein d'entrepôt. Une des caractéristiques majeures de nos travaux

est que nous ne souhaitions pas fixer à priori une structure logique et nous souhaitions que cet entrepôt autorise tout type d'interrogation pour faciliter les prises de décision (recherche d'information, analyse multidimensionnelle, interrogation assertionnelle). Les propositions d'Entrepôt de DOcument (EDO) étaient beaucoup trop restrictives et ne permettaient pas d'intégrer des documents sources hétérogènes plus ou moins structurés dans un entrepôt. Nous avons donc proposé un modèle d'EDO. Ce modèle permet de regrouper en classes les documents ayant une structure logique « proche ». Ce travail de modélisation a été accompagné par la définition d'un processus d'alimentation permettant d'extraire la structure logique du document, de comparer cette structure avec les structures contenues dans l'entrepôt et d'intégrer le nouveau document. Parmi ces trois phases, la phase de comparaison est la plus importante car elle est basée sur le calcul de similarité d'arborescences hétérogènes d'éléments ordonnés et étiquetés (représentant les structures logiques).

Présentation des données décisionnelles

Pour cette problématique, nous nous sommes concentrés sur la modélisation des magasins de données OLAP. Nous avons été confrontés au fait que la modélisation OLAP ne repose pas sur une formalisation standard, stable, reconnue par l'ensemble de la communauté [Rizzi et al., 2006]. En réponse, nous avons proposé un modèle conceptuel générique de base orienté décideur. Tout en faisant abstraction des aspects techniques, ce modèle permet de représenter l'ensemble des concepts présents dans les différentes propositions. Notamment, il permet de construire des schémas multi-faits avec des dimensions multi-hiérarchisées.

Seules, ces données brutes ne peuvent être suffisantes pour faciliter les prises de décision. Afin de répondre à de nouveaux besoins des décideurs, nous avons étendu ce modèle de base :

- Dans un premier temps, afin de pouvoir intégrer les données documentaires des entrepôts, nous avons proposé une typologie des dimensions en proposant notamment les dimensions documentaires représentant la structure logique des documents et en proposant des mesures non numériques avec des fonctions d'agrégation adaptée ;
- Dans un second temps, nous avons voulu répondre à un besoin crucial, à savoir disposer de données fiables permettant d'assurer des analyses décisionnelles pertinentes. Pour ce faire, nous avons proposé le concept de contrainte sémantique inter ou intra dimensions. Ces contraintes permettent de spécifier des contraintes d'exclusion, d'inclusion, de simultanéité entre les hiérarchies de dimensions afin d'interdire les corrélations de données incohérentes ;
- Dans un troisième temps, les décideurs souhaitaient manipuler des schémas multidimensionnels pouvant évoluer dans le temps. Dans ce cadre, nous avons proposé un schéma reposant sur le concept de version d'étoiles, de version de fait et de version de dimension. Cette modélisation présente l'avantage de pouvoir gérer les évolutions réelles de données tout en permettant des simulations ;
- Les données seules, ne peuvent suffirent pour prendre des décisions. Notamment, les décideurs souhaitent pouvoir intégrer des informations complémentaires à ces données et éventuellement partager ces données avec des collaborateurs. A notre connaissance, il n'y a pas de proposition permettant d'intégrer dans un même espace les données décisionnelles et l'expertise des décideurs. Pour répondre à ce besoin, nous avons proposé d'intégrer le concept d'annotations. Une annotation contient aussi bien des informations subjectives (contenu, type) que des informations objectives (identifiant, date de création, créateur, référence à une annotation père, point d'ancrage local ou global). Ces annotations peuvent être utilisées pour un usage personnel ou collectif ;

- Enfin, pour faciliter encore la tâche des décideurs, nous avons souhaité proposer une solution permettant de personnaliser les schémas multidimensionnels. Une seule proposition [Bellatreche et al., 2005] a travaillé avec un modèle cube dont les dimensions contient un seul attribut. Nous avons étendu cette proposition en proposant d'affecter un poids aux attributs et un langage ECA permettant de préciser le contexte de manipulation (opération suivant laquelle les priorités sont fixées) et le contexte d'utilisation (état selon lequel les priorités sont affectées).

Exploitation de données multidimensionnelles

Cette thématique visait à proposer des mécanismes orientés décideur pour manipuler les composants d'un schéma multidimensionnel. Nous avons proposé une algèbre regroupant l'ensemble des opérations existantes dans la littérature. Pour ce faire, nous avons défini un noyau minimal d'opérateurs unaires représentant les opérations effectuées par les décideurs (rotation, forage, sélection, ordonnancement, calcul d'agrégation, modification des indicateurs et des axes d'analyse). A l'instar de l'algèbre relationnelle, nous avons proposé des opérateurs avancés permettant de simplifier l'écriture des requêtes et accélérant les traitements. Enfin, nous avons proposé un ensemble d'opérateurs binaires permettant de combiner deux Tables Multidimensionnelles (TM) afin de faire l'union, la différence ou l'intersection de deux analyses. Or, notre objectif étant une orientation utilisateur, nous avons voulu proposer un langage graphique pour les analyses décisionnelles. A partir d'une représentation graphique d'un schéma conceptuel, ce langage graphique permet d'effectuer des requêtes de manière incrémentale et interactive. Ce langage présente trois avantages : (1) il présente le contenu d'un magasin multidimensionnel au travers d'un schéma conceptuel, (2) il permet de faire complètement abstraction des opérations associées aux manipulations graphiques et (3) il est complet au regard de l'algèbre que nous avons proposée précédemment. Afin de compléter nos travaux, nous avons proposé un langage assertionnel reposant sur une extension du langage SQL. Ce langage permet non seulement d'interroger les données multidimensionnelles (extension de la commande SELECT pour obtenir une TM à partir du contenu des faits et des dimensions) mais également de définir et contrôler ces données.

Démarche de conception de systèmes d'aide à la décision

Enfin, nous avons complété nos travaux en proposant une démarche de conception. Cette démarche mixte prend en compte les besoins utilisateurs et les données sources. Elle repose sur la description d'un ensemble de tâches et de règles permettant le passage d'une tâche à une autre. De plus, cette démarche propose d'étudier l'ensemble des besoins : tactiques, stratégique et système tout en fournissant des règles automatiques pour la définition de l'architecture d'un SAD. Enfin, afin de capitaliser l'expérience et la connaissance des concepteurs décisionnels, nous avons défini un catalogue de patrons permettant de préciser les différentes étapes, phases et tâches de notre démarche.

2 PERSPECTIVES GENERALES

Nous proposons trois axes d'études complémentaires à nos travaux : la modélisation des Systèmes d'Aide à la Décision (SAD), la gestion des données décisionnelles et l'analyse en ligne

Modélisation des systèmes d'aide à la décision

D'un point de vue méthodologique, nos propositions peuvent être complétées dans deux directions : l'analyse des besoins décisionnels et la méta-modélisation.

Dans un premier temps, nous avons défini les différentes étapes pour l'analyse, la conception et le développement de Systèmes d'Aide à la Décision (SAD) au travers d'une démarche mixte. L'étape d'analyse et notamment l'analyse des acteurs stratégiques doit être

développée. La compréhension et la traduction des besoins utilisateurs en des applications décisionnelles viables restent vitales pour l'avenir de l'organisation. De nos jours, l'analyse des besoins pour le développement d'applications transactionnelles est largement étudiée. Nous pouvons notamment citer la conférence internationale "Requierements engineering", dont la 15ème édition a eu lieu en octobre dernier. Dans ce cadre, mon objectif est de partager l'expérience des chercheurs de cette communauté avec nos compétences en conception de SAD. Quelques propositions ont déjà vu le jour avec l'adaptation de modèles de buts existants pour le décisionnel : GDI [Prakash & Gosain, 2003], i* [Giorgini et al., 2005 ; Mazon et al., 2005a] et MAP [Gam & Salinesi, 2006]. Nous souhaitons proposer une orientation plus globale et plus adaptée au travers :

- d'une analyse des besoins spécifiques (avec validation et vérification) basée sur des points de vue décisionnels exprimés au travers d'un langage adapté aux décideurs,
- d'une modélisation spécifique des acteurs, des domaines et des buts décisionnels avec un couplage de l'information, des sources de données et des traitements d'analyse exploratoire,
- de la transformation de cette analyse en un schéma conceptuel multidimensionnel.

A plus long terme, nous souhaitons créer un partenariat entre notre équipe et les personnes travaillant dans le domaine de l'intelligence économique au sein de notre Université. L'intelligence économique[26] (IE), définie comme "la maîtrise et la protection de l'information stratégique pertinente pour tout acteur économique", vise à offrir des grilles d'analyse d'enjeux liés à la compétitivité de l'économie et à la sécurité de l'État et des entreprises. Dans ce cadre, nous souhaitons proposer une méthode d'analyse et de conception spécifique à l'analyse des indicateurs de l'IE et à leur implantation dans un système d'aide à la décision.

Dans un second temps, nous souhaitons répondre à un besoin vital des SAD, à savoir la gestion des méta-données. En effet, par essence un SAD repose sur l'interconnexion d'outils répartis manipulés par des décideurs ayant des perspectives d'analyse différentes. Une première réponse a été donnée par les industriels avec l'OIM[27] (Open Information Model) du MDC (Meta-Data Coalition), aujourd'hui disparu, et le CWM[28] (Common Warehouse Model) de l'OMG. Le CWM vise à faciliter l'échange de méta-données entre les différents outils décisionnels dans un environnement réparti et hétérogène. Ces solutions se limitent à l'aspect technique et elles n'intègrent pas des méta-données permettant de représenter la gestion des droits utilisateurs, les visions personnalisées des décideurs, la qualité et la sécurité des données [Vetterli et al., 2000]. Dans le cadre des SAD, la méta-modélisation reste encore un champ d'investigation peu exploré. A titre d'exemple, dans la conférence DAWAK, nous trouvons un seul article relatif aux méta-données en 2007 [Farinha & Trigueiros, 2007] et aucun pour les éditions 2004, 2005 et 2006. Aussi, une perspective intéressante à nos travaux repose sur la proposition d'un méta-modèle complet, spécifique aux SAD. Cette réflexion doit aussi bien porter sur le contenu du méta-modèle que sur son implantation et sa gestion (centralisée, répartie ou mixte). Un tel méta-modèle doit aborder différents aspects :

- l'aspect technique : les méta-données techniques sont destinées aux développeurs et aux administrateurs afin d'implanter et de maintenir le SAD. Les méta-données techniques représentent des données descriptives (schémas et localisations des sources hétérogènes et réparties, de l'entrepôt de données et des magasins de données) et des données de transformation décrivant les processus ETL entre les différents espaces de stockage ;

[26] http://www.intelligence-economique.gouv.fr/rubrique.php3?id_rubrique=6
[27] http://xml.coverpages.org/mdc-oim.html
[28] http://www.cwmforum.org/ et http://www.omg.org/technology/documents/formal/cwm.htm

- l'aspect décisionnel : ces méta-données visent à faciliter la compréhension et l'exploitation des données du SAD par les décideurs. Ces données peuvent intégrer des modèles conceptuels, des modèles de l'entreprise avec la terminologie adaptée et la liaison entre ces termes et les données décisionnelles, des informations sur la qualité des données (provenance des données, fonction de transformation, périodicité et date de rafraîchissement..) [Do & Rahm, 2000] ;
- l'aspect utilisateur : ces méta-données permettent de préciser les profils, les rôles, les sujets d'intérêts, les données mais également les opérations d'analyses décisionnelles autorisées. Ces méta-données seront combinées avec celles relatives à la gestion des annotations, de la personnalisation, des cohérences temporelle et sémantique tel que présenté dans ce mémoire.

La gestion des données décisionnelles

Au niveau de la gestion des données, nous pouvons proposer deux axes de recherches : la confidentialité et la répartition des données décisionnelles.

Les données d'un SAD étant vitales pour l'avenir de l'organisation, il est nécessaire de mettre en œuvre, dès les phases d'analyse, la gestion de la confidentialité des données. De par l'architecture des SAD, il faut définir des politiques de sécurité pour les différents espaces de stockage et les différents traitements associés. Au niveau de l'Entrepôt de Données (ED), il faut préciser les données et les opérations accessibles par les utilisateurs (le plus souvent des administrateurs pour l'ED), les différents accès possibles sur les sources et les différentes extractions possibles pour la constitution des Magasins de Données (MD). Au niveau des MD, il faut déterminer pour chaque décideur, voire pour chaque groupe de décideurs, les données et les opérations d'analyse autorisées. Le point crucial de ces travaux est que cette gestion des droits ne se limite pas seulement, comme pour une BD, à l'accès aux données mais intègre également les droits d'accès aux opérations de l'analyse multidimensionnelle. Ces travaux compléteraient nos travaux relatifs aux contraintes dans les MD multidimensionnels [Ravat et al., 2005a], à la personnalisation de ce magasin [Ravat et al., 2007f] et à la définition d'une politique d'accès aux données multidimensionnelles [Sallami, 2004].

Les SAD sont appelés à gérer des volumes de données importants et les décideurs souhaitent une restitution rapide des données décisionnelles. Une première réponse a consisté à proposer une architecture décisionnelle basée sur la dualité ED – MD complétée par la sauvegarde des agrégations les plus usitées (vues matérialisées auxiliaires) afin d'accélérer les restitutions décisionnelles. Or, les données et les besoins d'un SAD évoluant constamment, il est nécessaire d'intégrer ces propositions dans une solution plus globale. Cette solution va combiner les propositions actuelles avec une répartition des données et des traitements. Pour l'ED, nous souhaitons offrir des mécanismes de fragmentation (horizontale, verticale ou mixte) qui tiennent compte des besoins des MD (modèle de données et politique de rafraîchissement) et des évolutions liées à son alimentation. Au niveau des MD, il est possible de combiner une hiérarchisation des magasins [Zhou et al., 2000] avec une fragmentation basée sur les données manipulées et les opérations d'analyse multidimensionnelle. Enfin, si nous souhaitons acheminer l'information pertinente aux décideurs, de manière non intrusive et sans que le décideur se soucie de la technologie de transmission de données, nous pouvons asseoir les principes précédemment définis sur une grille de calcul pervasive orientée utilisateur (comme défini dans [Pierson, 2005]). Une première réponse à cette problématique sera apportée par la réalisation du projet régional IAPA (Infrastructure d'Accès, de Partage et d'Analyse de données biomédicales), portée par le professeur JM Pierson et dont nous avons la responsabilité du lot 6 (analyse multidimensionnelle et décision).

Analyse en ligne

Au niveau de la manipulation des données, deux axes de recherche paraissent prometteurs : la définition d'une algèbre XML OLAP et la proposition d'un langage OLAP Mining.

A l'heure actuelle, il existe des propositions parcellaires mais il n'y a pas une véritable algèbre XML OLAP complète. Notamment, les auteurs de [Park et al., 2005] proposent des fonctions d'agrégation de mesures textuelles, les auteurs [Wang et al., 2003, 2005] proposent l'agrégation d'arbres XML. Cette algèbre XML OLAP vise à pouvoir exécuter des requêtes OLAP sur des données issues de sources XML. Pour répondre à ce besoin, il est nécessaire dans un premier temps de définir un modèle de données multidimensionnelles pour les documents XML. Dans un second temps, cette algèbre doit permettre d'appliquer les opérations multidimensionnelles telles que les forages et les rotations avec génération automatique et transparentes aux décideurs de commandes Xpath ou Xquery. Notamment, pour les forages, il faut définir des opérations d'agrégation de mesures textuelles telles que la sélection des X premiers mots clés, le résumé de textes et l'agrégation de mots clés. De plus, l'intégration des données semi-structurées implique la gestion de la relaxation d'arborescences XML [Amer-Yahia et al., 2002].

La dernière perspective que nous proposons est la définition d'un langage OLAP Mining. Un tel langage permettra de combiner les opérations de l'algèbre multidimensionnelle avec des fonctions de fouille de données (« datamining »). Quelques propositions ont été formulées pour appliquer la technique de classification et de prédiction sur des structures de données multidimensionnelles [Han, 1997 ; Messaoud et al., 2004 ; Zubcoff & Trujillo, 2006]. A l'heure actuelle, il n'y a pas eu la proposition d'un environnement d'aide à la décision permettant aux décideurs d'utiliser et de combiner des opérations d'analyse multidimensionnelles (telles que le forage ou les rotations) avec des fonctions ou des opérations de fouille de données.

BIBLIOGRAPHIE

Remarque : Toutes les références en gras sont celles dont je suis un des auteurs. Dans notre équipe, les auteurs sont classés par ordre alphabétique.

A

[Abelló et al., 2002] A. Abelló, J. Samos, F. Saltor "*YAM² (Yet Another Multidimensional Model): An Extension of UML*", International Database Engineering & Applications Symposium (IDEAS'02), Edmonton (Canada), p. 172-181, 17-19 Juillet 2002.

[Abelló et al., 2003] A. Abelló, J. Samos, F. Saltor, "*Implementing operations to navigate semantic star schemas*", 6th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2003), New Orleans (Louisiana, USA), p. 56–62, 7 Novembre 2003.

[Abelló et al., 2006] A. Abelló, J. Samos, F. Saltor, "*YAM²: a multidimensional conceptual model extending UML*", Information Systems (IS), vol.31(6), Elsevier, p. 541–567, Septembre 2006.

[Abiteboul et al., 2001] S. Abiteboul, V. Vianu, L. Segouphin "*Representing and Querying XML with Incomplete Information*", Symposium on Principles of Database Systems (PODS 2001), p. 150-161, Santa Barbara (California, USA), 21-23 Mai 2001.

[Abiteboul et al., 2002] S. Abiteboul, S. Cluet, G. Ferran, M.C. Rousset "*The Xyleme Project*", Computer Networks, 39(3), p. 225-238, Juin 2002.

[Abiteboul & Vianu, 1997] S. Abitboul, V. Vianu "*Queries and Computation on the Web*". 6th International Conference Database Theory (ICDT '97), Delphi, Greece, p. 262-275, 8-10 Janvier1997

[Abiteboul, 1997] S. Abitboul "*Querying semi-structured data*", 6th International Conference Database Theory (ICDT'97), Delphi, Greece, p. 1-18, 8-10 Janvier1997

[Aboud, 1990] M. Aboud "*Systèmes de recherche d'information : Thésaurus et classification*" Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), 1990.

[Adelberg, 1998] B. Adelberg "*NoDoSE - A Tool for Semi-Automatically Extracting Semi-Structured Data from Text Documents*" 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD'98), Seattle (Washington, USA), p. 283-294, 2-4 Juin 1998.

[Adler & Van Doren, 1972] M.J. Adler, C. Van Doren, "*How to Read a Book*", Simon & Shuster, 1972.

[Agrawal et al., 1995] R. Agrawal, A. Gupta, S. Sarawagi "*Modeling Multidimensional Databases*", IBM Research Report, http://rakesh.agrawal-family.com/papers/icde97olap_rj.pdf , 1995 .

[Agrawal et al., 1997] R. Agrawal, A. Gupta, S. Sarawagi "*Modeling Multidimensional Databases*", 13th International Conference on Data Engineering (ICDE'97), Birmingham (U.K.), p. 232–243, 7-11 April 1997.

[Agrawal et al., 2000] R. Agrawal, R. J. Bayardo Jr., R. Srikant, "*Athena: Mining-Based Interactive Management of Text Database*", 7th International Conference on Extending Database Technology (EDBT'00), Konstanz (Germany), LNCS 1777, Springer, p. 365–379, 27-31 March 2000.

[Alexander, 1977] C. Alexander, "*A Pattern Language: Towns, Buildings, Construction*", Oxford University Press, ISBN: 0195019199, 1977

[Amer-Yahia et al., 2002] S. Amer-Yahia, S. R. Cho, D. Srivastava, "*Tree Pattern Relaxation*", Advances in Database Technology, 8th International Conference on Extending Database Technology (EDBT'02), Prague (Czech Republic), LNCS 2287, p. 496–513, 25-27 March 2002.

[Annoni, 2003] E. Annoni, "*Conception et développement d'un langage assertionnel pour les bases de données multidimensionnelles*" Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2003.

[Annoni, 2007] E. Annoni "*Eléments méthodologiques pour le développement de systèmes décisionnels dans un contexte de reutilisation*", Thèse de doctorat en Informatique de l'Université des Sciences Sociales de Toulouse (Toulouse I), Juillet 2007.

[Annoni et al., 2005a] E. Annoni, F. Ravat, O. Teste, G. Zurfluh "*Une approche d'analyse et de conception de SID à base de patrons*", Revue des Sciences et Technologies de l'Information – Série RSTI – ISI (Ingénierie des Systèmes d'Information), Vol 10 - n°6/2005, J-P. Giraudin, D. Rieu, Hermès (Ed.), pp.81-106, ISBN 2-7462-1322-2, 2005.

[Annoni et al., 2005b] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, "*BIPAD : Une méthode d'analyse et de conception des systèmes d'information décisionnels par réutilisation de patron*", 10ème congrès de l'Association Information and Management (AIM'05), Toulouse, Septembre 2005.

[Annoni et al., 2006a] E. Annoni, F. Ravat, O. Teste, G. Zurfluh "*Towards Multidimensional Requirement Design*", 8th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'06), Krakow (Poland), Springer Verlag, LNCS 4081, p. 65-74, September 2006

[Annoni et al., 2006b] E. Annoni, F. Ravat, O. Teste, G. Zurfluh "*Automating the Choice of Decision Support Systems Architecture*", 17th International Conference on Databases and Expert Systems (DEXA'06), Krakow (Poland), Springer Verlag, LNCS 4080, p. 244-253, September 2006

[Annoni et al., 2006c] E. Annoni, F. Ravat, O. Teste "*Traitements à l'origine des systèmes d'information décisionnels*", Revue des Sciences et Technologies de l'Information – Série RSTI – ISI (Ingénierie des Systèmes d'Information), V. 11 n°6/2006, E. Métais, G. Zurfluh, Hermès (Ed.), p. 115-143, 2006

[Annoni et al., 2006d] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, "*Modélisation adaptée aux besoins utilisateurs dans le développement des systèmes d'information décisionnels*", Journées Entrepôts de Données et l'Analyse en ligne (EDA'06), Revue des Nouvelles Technologies de l'Information, RNTI-B-2, , Cépadues Edition, p. 23-38, Juin 2006.

[Annoni et al., 2006e] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, "*Méthode de Développement des Systèmes d'Information Décisionnels : Roue de Deming*", XXIVéme congrès INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID'06), Hammamet (Tunisie), p. 657-673, Juin 2006.

[Annoni et al., 2007] E. Annoni, F. Ravat, O. Teste, "*Data and Process analyses of Data Warehouse Requirements*", 19th Software Engineering and Knowledge Engineering (SEKE 2007), Boston (USA), July 09-11, 2007.

[Anton, 1987] JP Anton "*Contribution au développement des systèmes videotext multimedia*" Thèse d'état en Informatique de l'Université Paul Sabatier (Toulouse III), 1987

B

[Baeza-Yates & Ribeiro-Neto, 1999] R. Baeza-Yates, B. Ribeiro-Neto, "*Modern Information Retrieval*", Addison Wesley, ISBN 0-201-39829-X, 1999,

[Balmisse, 2002] G. Balmisse "*Gestion des connaissances : Outils et applications du knowledge management*", Edition Vuibert, Septembre 2002.

[Baralis et al., 1997] E. Baralis, S. Paraboschi, E. Teniente "*Materialized view selection in a multidimensional database*", 23rd International Conference on Very Large Data Bases (VLDB'97), Athens (Greece), p. 156-165, 25-29 August 1997.

[Barbier et al., 2004] F. Barbier, C. Cauvet, M. Oussalah, D. Rieu, S. Bennasri, C. Souveyet "*Concepts clés et techniques de réutilisation dans l'ingénierie des systèmes d'information*" Revue l'Objet, logiciel, bases de données, réseaux, Revue des Sciences et Technologies de l'Information, Hermes, Numéro spécial "Ingénierie des composants dans les systèmes d'information" sous la direction de M. Oussalah, D. Rieu, Vol. 10(1) pp.11-35, 2004.

[Baril, 1999] X. Baril, "*Historisation dans les entrepôts de données*" Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 1999.

[Bellahsene, 1998] Z. Bellahsene "*View Adaptation in Data Warehousing Systems*", 9th International Conference on Database and Expert Systems Applications (DEXA'98), Vienna (Austria), p. 300-309, 24-28 Août 1998.

[Bellatreche et al., 2005] L. Bellatreche, A. Giacometti, P. Marcel, H. Mouloudi, D. Laurent "*A personalization framework for OLAP queries*", 8th International Workshop on Data Warehousing and OLAP (DOLAP'05), Bremen (Germany), p. 9-18, November 4-5, 2005.

[Bellatreche et al., 2004] L. Bellatreche, M. Schneider, M. Mohania, H. Lorinquer, "*Bringing Together Partitioning, Materialized Views and Indexes to Optimize Performance of Relational Data Warehouses*" 6th International Conference on Data Warehousing and Knowledge Discovery (DaWak'04), Lecture Notes in Computer Science (LNCS) 3181, Zaragoza (Spain), p. 15-25, September 2004.

[Benakezou, 2006] L. Benakezou, "*Etude et Conception de bases de données multidimensionnelles temporelles*", Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2006.

[Benitez-Guerrero et al., 2003] E. Benitez-Guerrero, C. Collet, M. Adiba, "*Le système WHES pour l'évolution des entrepôts de données*", 19èmes Journées Bases de Données Avancées, (BDA '03), Lyon, 20-23 octobre 2003.

[Bertino et al., 1996] E. Bertino, E. Ferrari, G. Guerrini, "*A formal temporal object-oriented data model*" 5th International Conference on Extending Database Technology – (EDBT'96), p. 342-356, Avignon (France), LNCS 1057, p. 342-356, March 25-29, 1996.

[Bertino & Guerrini, 1998] E. Bertino, G. Guerrini "*Extending the ODMG Object Model with Composite Objects*" 13th ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages & Applications (OOPSLA '98), Vancouver (British Columbia, Canada), p. 259-270, October 18-22, 1998.

[Bhowmick et al., 2000a] S.S. Bhowmick, W. Keong Ng, S. Kumar Madria "*Web Schemas in WHOMEDA*" 3rd ACM International Workshop on Data Warehousing and OLAP (DOLAP 2000), Washington (USA), p. 17-24, 10 November 2000.

[Bhowmick et al., 2000b] S.S. Bhowmick, S. Madria, W. Keong Ng, E-P. lim "*Data visualisation operators for whowedia*", The computer Journal, Vol. 43(5), p.364-385, 2000.

[Blaschka et al., 1999] M. Blaschka, C. Sapia, G. Hoflin, "*On schema evolution in multidimensional databases*" 1st International Conference on Data Warehousing and Knowledge Discovery – (DaWaK'99), LNCS 1676, p. 153-164, Florence (Italy), August 30–Sept. 1, 1999.

[Body et al., 2002] M. Body, M. Miquel, Y. Bédard, A. Tchounikine, "*A multidimensional and multiversion structure for OLAP Applications*" 5th International Workshop on Data Warehousing and OLAP (DOLAP'02), McLean (Virginia,USA), p. 1-6, 8 November 2002.

[Boughanem, 2000] M. Boughanem, "*Formalisation et Spécification de Systèmes de Recherche et de Filtrage d'Information*" Habilitation à diriger des recherches en Informatique de l'Université Paul Sabatier (toulouse III), Novembre 2000.

[Bonifati et al., 2001] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, S. Paraboschi "*Designing data marts for data warehouses*" ACM Transactions on Software Engineering and Methodology, Vol. 10(4), p. 452-483, October 2001.

[Boucheikh, 2005] F. Boucheikh "*Méthodologie de conception de systèmes décisionnels*" Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Boussaïd et al., 2007a] O. Boussaïd, A. Tanasescu, F. Bentayeb, J. Darmont, "*Integration and dimensional modeling approaches for complex data warehousing*" Journal of Global Optimization, Vol. 37(4), p. 571-591, April 2007.

[Boussaïd et al., 2007b] O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher, "Warehousing complex data from the Web", International Journal of Web Engineering and Technology (IJWET), 2007 (à paraitre).

[Bouzeghoub & Kostadinov, 2005] M. Bouzeghoub, D. Kostadinov "*Personnalisation de l'information : aperçu de l'état de l'art et définition d'un modèle flexible de profils*", 2ème conférence francophone en Recherche d'Information et Applications (CORIA'05), Grenoble, pp. 201-218, 5-11 Mars 2005.

[Bouzguenda, 2002] L. Bouzguenda, "*Conception et implantation d'un prototype d'interrogation et de visualisation d'une base multidimensionnelle*", Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2002.

[Bret & Teste, 1999] F. Bret, O. Teste, "*Construction Graphique d'Entrepôts de données*" XVIIème Congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'99), La Garde (Var, France), p. 165-184, 02-04 Juin 1999.

[Bruckner et al., 2001a] R. M. Bruckner, B. List, J. Schiefer, A. Min Tjoa, "*Modeling Temporal Consistency in Data Warehouses*", 1st International Workshop on Knowledge Extraction for Enterprise Services (KEES) of 12th International Workshop on Database and Expert Systems Applications (DEXA Workshop), IEEE Computer Society, p. 901–905, 2001.

[Bruckner et al., 2001b] R. M. Bruckner, B. list, J. Schiefer "*Developing requirements for data warehouse systems with use cases*", Seventh Americas conference on Information Systems, Boston (Massachussets, USA), p. 329-335, August 3-5, 2001.

[Brush, 2002] A. J. B. Brush, "*Annotating Digital Documents for Asynchronous Collaboration*" Technical report 02-09-02, Department of Computer Science and Engineering, University of Washington, USA, 2002.

[Bukhres & Elmagarmid, 1995] O. A. Bukhres, A. K. Elmagarmid "*Object-Oriented Multidatabase Systems: A Solution for Advanced Applications*", Prentice Hall; 1st edition, ISBN-10: 0131038133, October 1995.

[Buschmann et al., 1996] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, M. Stal, "*Pattern-Oriented Software Architecture : A System of Patterns*",Wiley, 1 edition, ISBN-10: 0471958697, August 8, 1996.

[Buzydlowski et al., 1998] J. W. Buzydlowski, Il-Yeol Song, L. Hassell "*A Framework for Object-Oriented On-line Analytical Processing*" 1st International Workshop on Data Warehousing and OLAP (DOLAP '98), Bethesda (Maryland, USA), p.10-15, 7 November 1998.

C

[Cabanac et al., 2006a] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, "Modèle conceptuel pour bases de données multidimensionnelles annotées", Journées d'Extraction et de Gestion des Connaissances (EGC'06), Revue des Nouvelles Technologies de l'Information, RNTI-E-6, Vol. I, Cépadues (ed.), Lille, p.119-124, 17-20 Janvier 2006.

[Cabanac et al., 2006b] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, "*Méta-modélisation des bases de données multidimensionnelles annotées*", Journées Entrepôts de Données et l'Analyse en ligne (EDA'06), Revue des Nouvelles Technologies de l'Information, RNTI-B-2, Cépadues Edition, p. 39-54, Juin 2006.

[Cabanac et al., 2007] G. Cabanac, M. Chevalier, F. Ravat, O. Teste "*An Annotation Management System for Multidimensional Databases*", 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg (Germany), Springer-Verlag, LNCS 4654, p.89-98, 03-07 Septembre 2007.

[Cabibbo & Torlone, 1997] L. Cabibbo R. Torlone, "Querying Multidimensional Databases"*., 6th International Workshop* on *Database Programming Languages* (*DBPL-6*), LNCS 1369, Colorado, (USA), Springer, p. 319-335, August 18-20, 1997.

[Cabibbo & Torlone, 1998] L. Cabibbo, R. Torlone, "*From a Procedural to a Visual Query Language for OLAP*", 10th International. Conference on Scientific and Statistical Database Management (SSDBM'98), Capri (Italy), p. 74–83, July 1-3, 1998.

[Cabibbo & Torlone, 2000] L. Cabibbo, R. Torlone "*The Design and Development of a Logical System for OLAP*", 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000), LNCS 1874, London (UK), p. 1–10, September 4-6, 2000.

[Canillac 1991] M. Canillac, "*Un modèle et une interface hypertexte pour les bases de données orientées objet*", Thèse de doctorat de l'Université Paul Sabatier (Toulouse III), 11 Juillet 1991.

[Carneiro & Braymer, 2002] L. Carneiro, A. Brayner, "*X-META: A methodology for data warehouse design with metadata management*", 4th Intl. Workshop Design and Management of Data Warehouses (DMDW'02), Toronto (Canada), p. 13-22, 27 May 2002.

[Carpani & Ruggia, 2001] F. Carpani, R. Ruggia, "*An Integrity Constraints Language for a Conceptual Multidimensional Data Model*", 13th International Conference on Software Engineering & Knowledge Engineering (SEKE'01), Buenos Aires, Argentina, p. 220-227, June 13-15, 2001.

[Cattell, 1998] R.G.G. Cattell "*ODMG-93 : Standard des bases de données objet*", Vuibert (Ed.) ISBN : 2-8418-0006-7, 1998.

[Cavero et al., 2001] J. M. Cavero, M. Piattini, E. Marcos, "*MIDEA: A Multidimensional Data Warehouse Methodology*" 3rd International Conference on Enterprise Information Systems (ICEIS'01), Setubal (Portugal), Volume 1 p. 138-144, July 7-10, 2001.

[Chakrabarti et al., 1998] S. Chakrabarti, B. Dom, R. Agrawal, Prabhakar Raghavan, "*Scalable Feature Selection, Classification and Signature Generation for Organizing Large Text Databases into Hierarchical Topic Taxonomies*", The VLDB Journal, vol.7(3), p. 163–178, 1998.

[Chaudhuri & Dayal, 1997] Chaudhuri S., Dayal U., "*An Overview of Data Warehousing and OLAP Technology*", ACM SIGMOD Record, 26(1), pp 65-74, 1997.

[Chowdhury, 2004] G. G. Chowdhury, "Introduction to Modern Information Retrieval", 2nd edition, Facet Publishing, ISBN 1856044807, 2004.

[Chrisment et al., 2000] C. Chrisment, J. le Maître, F. Sèdes "*Bases de données documentaires*", Techniques de l'ingénieur, Traité Informatique H 7 248, Mai 2000.

[Chrisment et al., 2005] C. Chrisment, G. Pujolle, F. Ravat, O. Teste, G. Zurfluh, "*Les entrepôts de données*", Traités des Techniques de l'Ingénieur - H 3870, Technique de l'Ingénieur (ed.), Février 2005.

[Chrisment et al., 2006] C. Chrisment, G. Pujolle, F. Ravat, O. Teste, G. Zurfluh, "*Bases de données décisionnelles*" Encyclopédie de l'informatique et des systèmes d'information, Jacky Akoka, Isabelle Comyn-Wattiau (Eds.), Vuibert, I/5, p. 533-546, Novembre 2006.

[Christophidès, 1996] V. Christophides, "*Documents structures et bases de données objet*", Thèses de Doctorat du Conservatoire National des arts et Métiers (CNAM), Octobre 1996.

[Coad , 1992] P. Coad, "Object-Oriented Patterns" Communication of ACM Vol. 35(9), p. 152-159, September 1992.

[Codd et al., 1993] E. F. Codd, S. B. Codd, C.T. Salley, "*Providing OLAP to user analyst: an IT mandate*" , Rapport technique, E.F. Codd and associates, 1993.

[Comparot & Chrisment, 1994] C. Comparot-Poussier, C. Chrisment "*Hyperbase pour la gestion électronique de documents techniques*" Ingénierie des Systèmes d'Information, Vol. 2 (5), p. 533-570, Edition Hermes, 1994.

[Comparot, 1994] C. Comparot-Poussier "*HYPERBASE : formalisation et architecture*" Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), Janvier 1994.

[Conte et al., 2001] A. Conte, M. Fredj, J-P. Giraudin, D. Rieu, "*P-Sigma : un formalisme pour une représentation unifiée de patrons*" XIXème congrès INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID'01), Martigny (Suisse), p. 67-86, 24-27 mai 2001.

[Cui & Widom, 2000] Y. Cui, J. Widom "*Practical Lineage Tracing in Data Warehouses*" 16th International Conference on Data Engineering (ICDE'00), San Diego (California, USA), p. 367-378, 28 Février - 3 Mars, 2000.

D

[Datta & Thomas, 1999] A. Datta, H. Thomas, "*The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses*", Decision Support Systems (DSS), vol.27(3), Elsevier, p. 289–301, December 1999.

[Dayal, et al., 1988] U. Dayal, B. T. Blaustein, A. P. Buchmann, U. S. Chakravarthy, M. Hsu, R. Ledin, D. R. McCarthy, A. Rosenthal, S. K. Sarin, M. J. Carey, M Livny, R. Jauhari, "*The HiPAC Project: Combining Active Databases and Timing Constraints*", SIGMOD Record Vol. 17(1), p. 51-70, March 1988.

[Denjean, 1989] P. Denjean "*Interrogation d'un système Vidéotext arborescent : l'indexation automatique des textes*", Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), 1989.

[Do & Rahm, 2000] H.H. Do, E. Rahm, "*On Metadata interoperability in data warehouses*", Rapport technique 01-2000 (http://lips.informatik.uni-leipzig.de/pub/2000-13/), Mars 2000.

[Doucet et al., 1996] A. Doucet, S. Gançarski, G. Jomier, S. Monties, "*Integrity Constraints and Versions*", 6th International Workshop on Foundations of Models and Languages for Data and Objects (FMLDO'96), Dagstuhl, (Germany), p. 25-39, September 16-20, 1996.

E

[Eder & Koncilia, 2001] J. Eder, C. Koncilia, "*Changes of Dimension Data in Temporal Data Warehouses*", 3rd International Conference on Data Warehousing and Knowledge Discovery (DAWAK'01), Munich (Germany), LNCS 2114 p. 284-293, 2001.

[Evolution, 2001] Groupe Evolution "*Entrepôt de données pour l'aide à la décision*", Chapitre de l'ouvrage "Ingénierie des systèmes d'information", Traité IC2 série informatique et systèmes d'information, C. Cauvet et C. Rosenthal-Sabroux (Eds.), Février 2001.

F

[Farinha & Trigueiros, 2007] J. Farinha, M.J. Trigueiros "*An extensible metadata framework for the data quality assessment of composite structures*" 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg (Germany), Springer-Verlag, LNCS 4654, p. 34-44, September 03-07, 2007.

[Faulstich et al., 1997] L.C. Faulstich, M. Spiliopoulou, V. Linneman "*WIND : A warehouse for Internet data*", 15th British National Conference on Databases(BNCOD 1997), London (United Kindom), Lecture Notes in Computer Science 1271 Springer Verlag, p. 169-183, July 7-9, 1997.

[Faulstich et al., 1998] L.C. Faulstich, M. Spiliopoulou., "*Building HyperView Wrappers for Publisher Web-Sites*", European Conference on Digital Libraries (ECDL'98), Heraklion (Crete, Greece), Lecture Notes in Computer Science 1513 Springer, p. 115-134, September 21-23, 1998.

[Feki et al., 2006] J. Feki, H. Ben-Abdallah, M. Ben-Abdallah, "*Réutilisation des patrons en étoile*" XXIVéme congrès INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID'06), Hammamet (Tunisie), p. 687-701, Juin 2006.

[Fernandez, 2003] A.Fernandez, "*Les nouveaux tableaux de bord des managers - Le projet décisionnel dans sa totalité*", Editions d'Organisation, 3e édition, ISBN10 : 2-7081-3387-X, Janvier 2003

[Fondin, 1998] H. Fondin "*Le traitement numérique des documents*", Editions Hermes, 1998.

[Fowler, 2004] M. Fowler, "*UML 2.0*", traduction M-C. Baland et L. Carite, Campus Press, ISBN10: 2-7440-1713-2, Février 2004.

[Frakes & Yates, 1992] W.B. Frakes, R.B. Yates "*Information Retrieval Data Structures & Algorithms*", ISBN 0-134-63837-9, Addison Wesley Publishing Company, 1992.

[Franco & De Lignerolles, 2000] J-M. Franco, S. De Lignerolles "*Piloter l'entreprise grâce au data warehouse*" Éd. Eyrolles, ISBN 2-212-09146-X, 2000.

[Franconni & Kamble, 2004] E. Franconni, A. Kamble "*The GMD Data Model and Algebra for Multidimensional Information*", 16th International Conference on Advanced Information Systems Engineering (CaiSE'04), Lecture Notes in Computer Science 3084, Riga (Latvia), June 7-11, 2004.

[Fualdes, 2001] C. Fualdes "*Les entrepôts de documents : extraction et comparaison de structures logiques*", Rapport de DEA 2IL, Université Paul Sabatier (Toulouse III), Juin 2001.

[Fuhr, 2000] N. Fuhr "*Models in Information Retrieval*" 3th European Summer-School (ESSIR'00), Varenna (Italy), Lectures on Information Retrieval, p. 21-50, September 11-15, 2000.

G

[Gam & Salinesi, 2006] I. Gam, C. Salinesi, "*Un processus dirigé par les exigences pour la conception des entrepôts de données*" XXIVéme congrès INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID'06), Hammamet (Tunisie), p. 1023-1038, Juin 2006.

[Gamma et al., 1995] E. Gamma, R. Helm, R. Johnson, J. Vlissides, "*Design patterns, elements of reusable object-oriented software*", Addison-Wesley Professional Computing Series, 1st edition, ISBN-10: 0201633612, January 1995.

[Garcia-Molina et al., 1998] Garcia-Molina H., Labio W. J., Yang J., "*Expiring Data in a Warehouse*" 24rd International Conference on Very Large Data Bases (VLDB'98), New York City (New York, USA), p. 500-511, August 24-27, 1998.

[Gardarin & Yoon, 1996] G. Gardarin, S. Yoon "*HyWeb: Un Système d'Interrogation Orienté Objet pour le WEB*" 12èmes Journées Bases de Données Avancées (BDA'96), Cassis, p. 205-224, August 27-30, 1996.

[Gargouri, 2006] M. Gargouri, "Assistance à l'élaboration incrémentale d'un magasin de données" Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Ghozzi, 2000] F. Ghozzi, "*Mécanismes d'historisation et d'archivage pour les entrepôts de données complexes*" Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2000.

[Ghozzi, 2004] F. Ghozzi-Jedidi "*Conception et manipulation de bases de données dimensionnelles à contraintes*" Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), Novembre 2004.

[Ghozzi et al., 2003a] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, "*Constraints and Multidimensional Databases*", 5th International Conference on Enterprise Information Systems (ICEIS'03), Angers (France), pp.104-111, April 23-26, 2003.

[Ghozzi et al., 2003b] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, "*Modèle Multidimensionnel à Contraintes*", 3èmes Journées d'Extraction et de Gestion des Connaissances (EGC'03), Revue des Sciences et Technologies de l'Information, Série RIA-ECA (Extraction des Connaissances et Apprentissage), Volume 17 - n°1-2-3/2003, Lyon, - ISBN 2-7462-0631-5, p. 43-56, 22-24 Janvier 2003.

[Ghozzi et al., 2004] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh "*Contraintes pour modèle et langage multidimensionnels*", Revue des Sciences et Technologies de l'Information, Série RSTI - ISI (Ingénierie des Systèmes d'Information), Vol. 9, n°1/2004, Hermès (ed.), ISBN 2-7462-0914-4, Sélection des 6 meilleurs articles des 19ème Journées Bases de Données Avancées – BDA'03, p. 9-33, 2004.

[Ghozzi et al., 2005] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, "*Méthode de conception d'une base multidimensionnelle contrainte*", Journées Entrepôts de Données et l'Analyse en ligne (EDA'2005), Revue des Nouvelles Technologies de l'Information, RNTI-B-1, , Cépadues (ed.), p. 51-70, Juin 2005.

[Giorgini et al., 2005] P. Giorgini, S. Rizzi, M. Garzetti, "*Goal-oriented requirement analysis for data warehouse design*". 8th ACM International Workshop on Data Warehousing and OLAP (DOLAP'05), Bremen (Germany), p. 47-56, November 4–5, 2005.

[Goldman et al., 1999] R. Goldman, J. McHugh, J. Widom "*From Semistructured Data to XML: Migrating the Lore Data Model and Query Language*" ACM SIGMOD Workshop on The Web and Databases (WebDB'99), Philadelphia (Pennsylvania, USA), p. 25-30, June 3-4, 1999.

[Golfarelli et al., 1998] M. Golfarelli, D. Maio, S. Rizzi, "*The Dimensional Fact Model: A Conceptual Model for Data Warehouses* ", invited paper, Intlernational Journal of Cooperative Information Systems (IJCIS), vol.7(2-3), p. 215–247, 1998.

[Golfarelli & Rizzi, 1998] M. Golfarelli, S. Rizzi "*Methodological Framework for Data Warehouse Design*" 1st International Workshop on Data Warehousing and OLAP (DOLAP'98), Bethesda (Maryland, USA), p. 3-9, November 7, 1998.

[Gray et al., 1996] J. Gray, A. Bosworth, A. Layman, H. Pirahesh, "*Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total*". 12th International Conference on Data Engineering (ICDE'96), New Orleans (Louisiana), pp. 152–159, February 26 - March 1, 1996.

[Gupta & Mumick, 1995] A. Gupta, I.S. Mumick, "*Maintenance of Materialized Views: Problems, Techniques, and Applications*", IEEE Bulletin on Data Engineering; Special Issue on Materialized Views and Data Warehousing, 18 (2), p. 3-18, June 1995.

[Gupta et al., 1997] H. Gupta, V. Harinarayan, A. Rajaraman, J. D. Ullman "*Index Selection for OLAP*" 13th International Conference on Data Engineering (ICDE'97), Birmingham (U.K.), p. 208-219, April 7-11, 1997.

[Gupta & Mumick, 1999] H. Gupta, I.S. Mumick "*Selection of Views to Materialize Under a Maintenace-Time Constraint*", 7th International Conference on Database Theory – (ICDT '99), Jerusalem (Israel), p. 453-470, January 10-12, 1999.

[Gupta, 1997] H. Gupta "*Selection of Views to Materialize in a Data Warehouse*", , 6th International Conference on Database Theory (ICDT '97), Delphi (Greece), p. 98-112, January 8-10, 1997.

[Gyssens & Lakshmanan, 1997] M. Gyssens, L. V. S. Lakshmanan, "*A Foundation for Multi-dimensional Databases*", 23rd International. Conference on Very Large Data Bases (VLDB'97), Athens (Greece), p. 106–115, August 25-29, 1997.

[Gyssens et al., 1996] M. Gyssens, L. V. S. Lakshmanan, I. N. Subramanian, "*Tables as a Paradigm for Querying and Restructuring*", 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'96), Montreal (Canada), p. 93–103, June 3-5, 1996.

H

[Han, 1997] J. Han, "OLAP Mining: an integration of OLAP with Data Mining", 7th Conference on Database Semantics (DS-7), Leysin (Switzerland), p. 3-20, October 7-10, 1997.

[Hahn et al., 2000] K. Hahn, C. Sapia, M. Blaschka, "*Automatically Generating OLAP Schemata from Conceptual Graphical Models*", 3rd ACM International Workshop on Data Warehousing and OLAP (DOLAP 2000), Washington (USA), p. 9–16, November 10, 2000.

[Harinarayan et al., 1996] V. Harinarayan, A. Rajaraman, J. Ullman "*Implementing Data Cubes Efficiently*", 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD'96), Montreal, Quebec, Canada, p. 205-216, June 4-6, 1996.

[Horner et al., 2004] J. Horner, I-Y. Song, P.P. Chen "*An analysis of additivity in OLAP systems*", 7th ACM International Workshop on Data Warehousing and OLAP (DOLAP'04), Washington (DC, USA), pp. 83-91, November 12-13, 2004.

[Hull & Zhou, 1996] R. Hull, G. Zhou "*A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches*", 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD'96), Montreal, Quebec, Canada, p. 481-492, June 4-6, 1996.

[Hurtado & Mendelzon, 2002] C. A. Hurtado, A. O. Mendelzon "*OLAP Dimension Constraints*" 21th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'02), Madison (Wisconsin, USA), p. 169-179, June 3-5, 2002.

[Hurtado et al, 1999] C. A. Hurtado, A. O. Mendelzon, A. A. Vaisman "*Maintaining Data Cubes under Dimension Updates*" 15th International Conference on Data Engineering (ICDE'99), Sydney (Austrialia), p. 346-355, March 23-26, 1999.

[Husemann et al., 2000] B. Husemann, J. Lechtenborger, G. Vossen "*Conceptual data warehouse modelling*" 2nd International Workshop on Design and Management of Data Warehouses (DMDW'00), p. 6.1 -- 6.11, Stockholm (Sweden), June 5-6, 2000.

[Huyn, 1996] [Huyn 1996] Huyn N., "*Efficient View Self-Maintenance*", In Proceedings of the ACM Workshop on Materialized Views: Techniques and Applications (VIEW 1996), Monteal, Canada, Friday, p. 17-25, June 7, 1996.

[Huyn, 1997] Huyn N., "*Multiple-View Self-Maintenance in Data Warehousing Environments*", 23rd International Conference on Very Large Data Bases (VLDB'97), Athens (Greece), p. 26-35, August 25-29, 1997.

I

[Inmon, 1996] W. H. Inmon, "*Building the Data Warehouse*", John Wiley and Sons, New York, NY, deuxième edition, ISBN : 04771-14161-5, 1996.

[ISO, 1986] International Standard ISO 8879, "*Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML)*", (http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387) 1986.

J

[Jarke & Vassiliou, 1997] M. Jarke, Y. Vassiliou, "*Data Warehouse Quality: A Review of the DWQ Project*", Second Conference on Information Quality (IQ'97), Massachussets Institut of technology, Cambridge, p. 299-313, 1997.

[Jarraya, 2001] T. Jarraya, "*Conception et implantation des mécanismes d'extraction des données sources pour construire un entrepôt de données*", Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2001.

[Jerbi, 2007] H. Jerbi, " *Mémoire d'expertises décisionnelles à base d'annotations*", Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Jones & Song, 2005] M. E. Jones, I-Y. Song, "Dimensional modeling: identifying, classifying & applying patterns" 8th International Workshop on Data Warehousing and OLAP (DOLAP'05), Bremen (Germany), p. 29-38, November 4-5, 2005.

K

[Kaddes, 2005] M. Kaddes, "Etude de faisabilité d'une modélisation en constellation sous contraintes sémantiques", Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Kahan et al., 2002] J. Kahan, M-R. Koivunen, E. Prud'Hommeaux, R.R.Swick, "Annotea: an open RDF infrastructure for shared Web annotations" Computer Networks, Vol. 2(5), pp.589-608, 2002.

[Kedad & Métais, 1999] Z. Kedad, E. Métais "*Dealing with Semantic Heterogeneity During Data Integration*" 18th International Conference on Conceptual Modeling (ER '99), Paris (France), p. 325-339, November 15-18, 1999.

[Keith et al., 2005] S. Keith, O. Kaser, D. Lemire "*Analyzing Large Collections of Electronic Text Using OLAP*", 29th Annual Conference on Mathematics, Statistics and Computer Science (APICS 2005), Wolville (Canada); October 21-23, 2005.

[Khrouf, 2001c] K. Khrouf "*Design of Textual Dataweb*" 3rd International Conference on Enterprise Information Systems (ICEIS'01), Setúbal (Portugal), Vol. 1, p. 239-243, July 07-10, 2001.

[Khrouf, 2004] K. Khrouf "*Entrepôts de documents : de l'alimentation à l'exploitation*" Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), Juillet 2004

[Khrouf & Soulé-Dupuy, 2003] K. Khrouf, C. Soulé-Dupuy "*Vers une mémoire d'entreprise via les entrepôts de documents : extraction de structures logiques*" Troisièmes journées Extraction et Gestion des Connaissances (EGC 2003), Lyon (France),. Revue des Sciences et Technologies de l'Information - série RIA ECA 17(1-3) p. 201-206, 22-24 janvier 2003.

[Khrouf & Soulé-Dupuy, 2004] K. Khrouf, C. Soulé-Dupuy "*A Textual Warehouse Approach: A Web Data Repository*" Intelligent Agents for Data Mining and Information Retrieval. Idea Publishing Group, ISBN : 1-59140-277-8, p. 101-124, 2004.

[Khrouf et al., 2003] K. Khrouf, F. Ravat, C. Soulé-Dupuy "*Comparaison et fusion de structures logiques de documents semi-structurés*", Revue des Sciences et Technologies de l'Information, Série RSTI - ISI (Ingénierie des Systèmes d'Information), Vol. 8, n°5-6/2003, Hermès (ed.), ISBN 2-7462-0865-2, pp. 127-151, 2003.

[Khrouf at al., 2007a] K. Khrouf, M. Mbarki, F. Ravat, C. Soulé-Dupuy, N. Vallès-Parlangeau, "*Entreposage de documents multimédia : modélisation basée sur le contenu et instanciation automatique*", Revue CID Centre de Hautes études Internationales d'Informatique Documentaire, (à paraître)

[Khrouf at al., 2007b] K. Khrouf, M. Mbarki, F. Ravat, C. Soulé-Dupuy, N. Vallés-Parlangeau "*Les entrepôts de documents : Gestion des versions*", Colloque Veille Stratégique Scientifique et Technologique (VSST'07), Marrakech (Maroc), , IADIS Digital Library, 21-25 Octobre 2007.

[Kimball & Ross, 2002] R. Kimball, M. Ross, "*The Data Warehouse Toolkit : the complete guide to dimensional modelling"*, (deuxième edition), Wiley, 2002.

[Kimball et al., 2005] R. Kimball, L. Reeves, M. Ross, W. Thornthwaite "*Le data warehouse : Guide de conduite de projet*" (deuxième tirage), ISBN 2-212-11600-4, 2005.

[Korfhage, 1997] R. Korfhage "*Information storage and retrieval*", Wiley Computer Publishing, ISBN 0-471-14-338, 1997.

[Kotidis & Roussopoulos 1999] Y. Kotidis, N. Roussopoulos "*DynaMat: A Dynamic View Management System for Data Warehouses*", , 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD 1999), Philadelphia, (Pennsylvania, USA), p. 371-382, June 1-3, 1999.

[Koussa, 2007] C. Koussa, "Bases de données multidimensionnelles : construction d'un magasin de données à partir de sources XML", Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Koutrika & Ioannidis, 2004] G. Koutrika, Y. Ioannidis "P*ersonalization of queries in database systems*", 20th International conference on Data Engineering (ICDE'04), Boston, USA, pp597-608, 30 March - 2 April 2004.

L

[Labio & Garcia-Molina, 1996] W. J. Labio, H. Garcia-Molina "*Efficient Snapshot Differential Algorithms for Data Warehousing*", 22th International Conference on Very Large Data Bases, (VLDB'96), Mumbai (Bombay), India, p. 63-74, September 3-6, 1996.

[Labio et al., 1997] W. J. Labio, D. Quass, B. Adelberg "*Physical Database Design for Data Warehousing*", 13th International Conference on Data Engineering (ICDE'97), Birmingham (U.K.), p. 277-288, April 7-11 1997.

[Labio et al., 1999] W. J. Labio, R. Yerneni, H. Garcia-Molina "*Shrinking the Warehouse Update Window*", 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD'99), Philadephia, Pennsylvania, USA, p. 383-394, June 1-3, 1999.

[Lallich & Ouerfelli, 1998] G. Lallich, T. Ouerfelli "*La segmentation pour l'indexation d'un document technique : Principe et méthodes*", Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatique, Sfax, Tunisie, Novembre 1998.

[Lamping & Rao, 1994] J. Lamping, R. Rao, "Laying Out and Visualizing Large Trees Using a Hyperbolic Space" 7th ACM Symposium on User Interface Software and Technology, Marina del Rey (CA, USA), p.13-14, November 2-4, 1994,

[Lapujade & Ravat, 1997] A. Lapujade, F. Ravat "*Conception de systèmes d'information multimédia répartie : application au milieu hospitalier*" XVème congrès Congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'97), p. 163-184, Toulouse (France), 10-13 Juin 1997.

[Laurent et al., 2002] A. Laurent, P. Marcel, F. Ravat, O. Teste, G. Zurfluh, "*Entrepôts de données et OLAP : un aperçu orienté recherche*", Rapport Final du groupe de travail GafOLAP – Action Spécifique n°20 « Extraction et Fouille » du CNRS-STIC, Juin 2002.

[Lechtenbörger & Vossen, 2003] J. Lechtenbörger, G. Vossen, "Multidimensional normal forms for data warehouse design" Information Systems Vol. 28(5), p. 415-434, Juillet 2003.

[Lehner, 1998] W. Lehner, "*Modelling Large Scale OLAP Scenarios*", 6th International Conference on Extending Database Technology - Advances in Database Technology (EDBT'98), LNCS 1377, Springer, p. 153–167, 1998.

[Le Parc 1997] A. Le Parc, "*Une algèbre et un langage graphique pour les bases de données objet intégrant le concept de versions*", Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), Décembre 1997.

[Le Thi, 2003] B. K. Le Thi, "Intégration de contraintes dans OLAP-SQL" Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2003.

[Li & Wang, 1996] C. Li, X. S. Wang, "*A Data Model for Supporting On-Line Analytical Processing*", 5th International. Conference on Information and Knowledge Management (CIKM'96), Rockville (Maryland, USA), p. 81–88, November 12 - 16, 1996.

[Lim & Kim, 2004] Y. Lim, M. Kim, "*A Bitmap Index for Multidimensional Data Cubes*" 15th International Conference on Database and Expert Systems Applications (DEXA'04), Zaragoza (Spain), p. 349-358, August 30-September 3, 2004.

[List et al., 2002] B. List, R. M. Bruckner, K. Machaczek, J. Schiefer, "*A comparison of data warehouse development methodologies, case study of the process warehouse*", 13th International Conference on Database and Expert Systems Applications (DEXA 2002), Aix-en-Provence - France, Lecture Notes in Computer Science N°2453, Springer Verlag, ISBN 3-540-44126-3, p. 203-215, September 2-6, 2002.

[Luján-Mora & Trujillo, 2003] S. Luján-Mora, J. Trujillo, "*A Comprehensive Method for Data Warehouse Design*". 5th International Workshop on Design and Management of Data Warehouses (DMDW'03) Berlin (Germany), September 8, 2003.

[Luján-Mora et al., 2004] S. Luján-Mora, J. Trujillo, P. Vassiliadis "*Advantages of UML for Multidimensional Modeling*", 6th International Conference on Enterprise Information Systems (ICEIS'04), Volume I - Databases and Information Systems Integration, Porto (Portugal), p. 298-305, April 14-17, 2004.

[Luján-Mora et al., 2006] S. Luján-Mora, J. Trujillo, I. Song, "*A UML profile for multidimensional modeling in data warehouses*", Data & Knowledge Engineering (DKE), vol.59(3), Elsevier, p. 725–769, December 2006.

M

[Malinowski & Zimányi, 2006] E. Malinowski, E. Zimányi "*Hierarchies in a multidimensional model: From conceptual modeling to logical representation*", Journal of Data & Knowledge Engineering, Volume 59(2), p. 348-377, November 2006.

[Marcel, 1998] P. Marcel "*Manipulation de données multidimensionnelles et langages de règles*", Thèse de doctorat de l'Institut des Sciences Appliquées de Lyon, 1998.

[Marshall, 1998] C. Marshall, "Toward an ecology of hypertext annotation", 9ème ACM conference on Hypertext and hypermedia: links, objects, time and space, Pittsburgh (USA),p. 40-49, 1998

[Mazon et al., 2005a] J-N. Mazon, J. Trujillo, M. Serrano, M. Piattini, "*Designing Data Warehouses: From Business Requirement Analysis to Mutidimensional Modelling*" 1st International Workshop on Requirements Engineering for Business Need and IT Alignment (REBNITA'05), Paris, Paper 7, August 29-30, 2005.

[Mazon et al., 2005b] J-N. Mazon, J. Trujillo, M. Serrano, M. Piattini "Applying MDA to the Development of Data Warehouses" 8th ACM International Workshop on Data Warehousing and OLAP (DOLAP'05), p. 57-66, Bremen (Germany), November 4–5, 2005.

[McCabe et al., 2000] C. McCabe, J. Lee, A. Chowdhury, D. A. Grossman, O. Frieder "*On the design and evaluation of a multi-dimensional approach to information retrieval*", 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR 2000), Athens (Greece), pp. 363-365, July 24-28 2000.

[McHugh et al., 1997] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, J. Widom "*Lore: A Database Management System for Semistructured Data*" SIGMOD Record 26(3): p. 54-66, September 1997.

[Mélèse, 1972] J. Mélèse "*L'analyse modulaire des systèmes de gestion, AMS*", Editions Hommes et Techniques, Paris, 1972.

[Mendelzon & Vaisman, 2000] A. O. Mendelzon, A. A. Vaisman "*Temporal Queries in OLAP*" 26th International Conference on Very Large Data Bases (VLDB'00), Cairo (Egypt), p. 242-253, 10-September 14, 2000.

[Mendelzon et al, 2003] A. O. Mendelzon, A. A. Vaisman "*Time in multidimensional databases*", Chapitre VI, Multidimensional Databases: Problems and Solutions, Maurizio Rafanelli (Ed.), Idea Publishing Group (IGP), ISBN 1-59140-053-8, p. 166–199, 2003.

[Messaoud et al., 2004] R. B. Messaoud, O. Boussaid, S. Rabaséda, "A new OLAP aggregation based on the AHC technique", 7th ACM International Workshop on Data Warehousing and OLAP (DOLAP'04), Washington DC (USA), p. 65–72, November 12-13, 2004.

[Miller, 1995] G.A. Miller "*WorldNet: A Lexical Database for English*", Communications of the ACM, Vol. 38(11), p. 39-41, November 1995.

[Moody & Kortink 2000] D. L. Moody, M. A. R. Kortink "*From enterprise models to dimensional models: a methodology for data warehouse and data mart design*", 2nd International workshop on Design and Management of Data Warehouses (DMDW'00), Stockholm (Sweden), Paper 5, June 5-6, 2000.

[Mothe, 2000] J. Mothe, "*Recherche et exploration d'informations -Découverte de connaissances pour l'accès; à l'information*" Habilitation à diriger des recherches en Informatique de l'Université Paul Sabatier (Toulouse III), Décembre 2000.

[Mothe et al., 2000] J. Mothe, F. Ravat, F. Riahi, G. Zurfluh. "*Structuration and enrichment of HTML documents in order to build a specific information warehouse*" 8th European Conference on Information System (ECIS'00), Vienne (Austria), p. 386-395, July 3-5, 2000.

[Mothe et al., 2003] J. Mothe, C. Chrisment, B. Dousset, J. Alau "*DocCube: "Multi-dimensional visualisation and exploration of large document sets*", Journal of the American Society for Information Science and Technology (JASIST), vol.54(7), pp. 650-659, May 2003.

[Mumick et al., 1997] I. Mumick, D. Quass, B. Mumick "*Maintenance of Data Cubes and Summary Tables in a Warehouse*", 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD'97), Tuscon (Arizona, USA), p. 100-111, May 13-15, 1997.

[Munzner, 2000] T. Munzner "Interactive visualization of large graphs and networks", Ph D. Dissertation, Stanford University, Juin 2000.

N

[Nanci & Espinasse, 2001] D. Nanci, B. Espinasse "*Ingénierie des systèmes d'information : Merise, Deuxième génération*", quatrième édition, Vuibert, ISBN 2-7117-8674-9, 2001.

[Negre, 2005] E. Negre, "Evolution de schémas dans une constellation" Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Nestorov et al., 1998] S. Nestorov, S. Abiteboul, R. Motwani "*Extracting Schema from Semistructured Data*" 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD'98), Seattle (Washington, USA), p. 295-306, June 2-4, 1998.

[Nguyen et al., 2000] T. B. Nguyen, A. Min Tjoa, R. Wagner, "*An Object Oriented Multidimensional Data Model for OLAP* ", 1st International Conference on Web-Age Information Management (WAIM 2000), Shanghai (China), LNCS 1846, p. 69-82, June 21-23, 2000.

[Niemi et al., 2003] T. Niemi, L. Hirvonen, K. Järvelin "*Multidimensional Data Model and Query Language for Informetrics*", Journal of the American Society for Information Science (JASIST), Vol. 54(10), p. 939-951, 2003.

O

[Ouerfelli, 2000] T. Ouerfelli "*Recherche d'information dans un document technique : Restructuration du texte dans une perspective de consultation sur indices textuels*", Thèse de doctorat en Informatique de l'Université Stendhal (Grenoble III), 2000.

P

[Parent & Spaccapietra, 1996] C. Parent, S. Spaccapietra "*Intégration de bases de données : panorama des problèmes et des approches*" Ingénierie des systèmes d'information, Vol. 4 (3), 1996.

[Park et al., 2005] B-K. Park, H. Han, I-Y. Song "*XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses*", 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Copenhagen (Denmark), LNCS 3589, Springer, pp.32-42, August 22-26, 2005.

[Pedersen & Jensen, 1998] T. B. Pedersen, C. S. Jensen "*Research Issues in Clinical Data Warehousing*" 10th International Conference on Scientific and Statistical Database Management (SSDBM'98), Capri (Italy), p. 43-52, July 1-3, 1998.

[Pedersen & Jensen, 1999] T. B. Pedersen, C. S. Jensen "*Multidimensional Data Modeling for Complex Data*", 15th International Conference on Data Engineering (ICDE'99), Sydney (Austrialia), p. 336-345, 23-26 March 1999.

[Pedersen et al., 2001] T. B. Pedersen, C. S. Jensen, C. E. Dyreson, "*A foundation for capturing and querying complex multidimensional data*", Information Systems, vol. 26(5), p. 383–423, July 2001.

[Pendse, 2005] N. Pendse, "*The OLAP Report :What is OLAP?",* Version du 15 Août 2005 (http://www.olapreport.com/fasmi.htm).

[Pendse, 2006] Ni. Pendse "*The OLAP Report : The origins of today's OLAP products*", Version du 17 Avril 2006, (http://www.olapreport.com/origins.htm).

[Pierson, 2005] JM Pierson, "*Une grille pervasive vue du côté des données*", mémoire d'habilitation à diriger des recherches en Informatique de l'Institut National des Sciences Appliquées de Lyon, 17 Novembre 2005

[Phipps & Davis, 2002] C. Phipps, K. C. Davis "*Automating data warehouse conceptual schema design and evaluation*", 4th Intl. Workshop Design and Management of Data Warehouses (DMDW'02), Toronto (Canada), p. 23-32, May 27, 2002.

[Prakash & Gosain, 2003] N. Prakash, A. Gosain, "*Requirements Driven Data Warehouse Development*", 15th Conference on Advanced Information Systems Engineering (CAiSE'03), Klagenfurt (Velden, Austria), Short Paper Proceedings p. 13-16, June16-20, 2003.

[Prat et al., 2006] N. Prat, J. Akoka, I. Comyn-Wattiau, "*A UML-based data warehouse design method*", Decision Support System, Vol.42(3), p. 1449-1473, 2006.

[Prat & Akoka, 2002] N. Prat, J. Akoka, "*From UML to ROLAP Multidimensional Databases using a Pivot Mode*l" 18èmes Journées Bases de Données Avancées (BDA '02), Evry, 21-25 octobre 2002.

Q

[Quass & Widom, 1997] D. Quass, J. Widom "*On-Line Warehouse View Maintenance for Batch Updates*", 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD'97), Tuscon (Arizona, USA), p. 147-158, May 13-15, 1997.

[Quass et al., 1996] D. Quass, A. Gupta, I. Mumick, J. Widom "*Making Views Self-Maintainable for Data Warehousing*", 4th Conference on Parallel and Distributed Information Systems (PDIS'96), Miami Beach (Florida, USA), p. 158-169, December 1996.

R

[Radev et al., 2002] D. R. Radev, E. H. Hovy, K. McKeown "*Introduction to the Special Issue on Summarization*" Computational Linguistics, Vol. 28(4), p. 399-408, Décembre 2002.

[Rafanelli, 2003] M. Rafanelli, "*Operators for Multidimensional Aggregate Data*". Chapter 5 of Multidimensional Databases: Problems and Solutions, M. Rafanelli (ed.), Idea Group Publishing, ISBN 1-59140-053-8, pp. 116-165, 2003.

[Ravat & Teste, 2000a] F. Ravat, O. Teste "*Object-Oriented Decision Support System*", Contribution à l'ouvrage « Enterprise Information Systems II », Kluwer Academic Publishers - ISBN 0-7923-7177, sélection des meilleurs articles de la 2nd International Conference on Enterprise Information Systems (ICEIS'00), p. 42-48, July 2001.

[Ravat & Teste, 2000b] F. Ravat, O. Teste "*A Temporal Object-Oriented Data Warehouse Model*" 11th International Conference on Database and Expert Systems – (DEXA'00), London (UK), Springer Verlag, Lecture Notes in Computer Science N°1873 p. 583-592, September 2000.

[Ravat & Teste, 2000c] F. Ravat, O. Teste "*An Object Data Warehousing Approach: a Web Site Repository*", 4th East-European Conference on Advances in Databases and Information Systems – DASFAA-ADBIS'00, Prague (Czech Republic), ISBN 80-85863-56-1, p. 128- 137, September 5-8, 2000.

[Ravat & Teste, 2001] F. Ravat, O. Teste, "*Modélisation et manipulation de données historisées et archivées dans un entrepôt orienté objet*", 17ième Journées Bases de Données Avancées (BDA'01), Cépadues Editions, Agadir (Maroc), ISBN 2-85428-570-0, p. 243-256, 29 Octobre - 2 Novembre 2001.

[Ravat & Teste, 2006] F. Ravat, O. Teste "*Supporting Data Changes in Multidimensional Data Warehouses*", International Review on Computers and Software, Praize Worthy Prize, Wantag - USA, Vol. 1(3), p. 251-259, November 2006.

[Ravat et al., 1997] F. Ravat, M. De Michiel, G. Zurfluh "*Distributed Object Oriented Databases: An Allocation Method*". 8th International Conference on Database and Expert Systems Applications (DEXA'97), Toulouse (France), Lecture Notes in Computer Science N°1308, Springer, p. 367-376, ISBN 3-540-63478-9, September 1-5, 1997.

[Ravat et al., 1999] F. Ravat, O. Teste, G. Zurfluh "*Towards Data Warehouse Design*", 8th International Conference on Information and Knowledge Management – (CIKM'99), Kansas City (Missouri, USA), ACM Press - ISBN 1-58113-146-1, p. 359-366, November 2-6 1999.

[Ravat et al., 2000] F. Ravat, O. Teste, G. Zurfluh, "*Modélisation et extraction de données pour un entrepôt objet*", 16ième Journées Bases de Données Avancées (BDA'00), Blois (France), p 119-138, 24-27 Octobre 2000.

[Ravat et al., 2001] F. Ravat, O. Teste, G. Zurfluh, "*Modélisation multidimensionnelle des systèmes décisionnels*", 1ères Journées d'Extraction et de Gestion des Connaissances (EGC'01),Revue des Sciences et Technologies de l'Information, Série RIA-ECA (Extraction des Connaissances et Apprentissage), Nantes, Volume 1 - n°1-2/2001 - ISBN 2-7462-0216-6, p. 201-212, 17-19 Janvier 2001.

[Ravat et al., 2002] F. Ravat, O. Teste; G. Zurfluh "*Langages pour Bases Multidimensionnelles : OLAP-SQL*", Revue des Sciences et Technologies de l'Information, série ISI-NIS (Ingénierie des Systèmes d'Information), Vol. 7, n°3/2002, Hermès (ed.), - ISBN 2-7462-0579-3, pp.11-38, 2002.

[Ravat et al., 2005a] F. Ravat, O. Teste, G. Zurfluh "*Constraint-Based Multi-Dimensionnal Databases*", Chapitre XI de l'ouvrage "Database Modeling for Industrial Data Management", sous la direction de Zongmin Ma, IDEA Group (ed.) - ISBN 1-59140-685-4, (article envoyé en Octobre 2004 au comité de programme international, revisité en Avril 2005 et accepté en Octobre 2005), p. 323-368, 2006.

[Ravat et al., 2005b] F. Ravat, O. Teste, G. Zurfluh, "*Manipulation et fusion de données multidimensionnelles*", 5èmes Journées d'Extraction et de Gestion des Connaissances (EGC'05), Revue des Nouvelles Technologies de l'Information, RNTI-E-3, Vol. I, Cépadues (ed.), Paris, p. 349-354, 19-21 Janvier 2005.

[Ravat et al., 2006a] F. Ravat, O. Teste, G. Zurfluh "*A Multiversion-based Multidimensional Model*" 8th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'06), Krakow (Poland), Springer Verlag, LNCS 4081, pp. 75-84, September 2006.

[Ravat et al., 2006b] F. Ravat, O. Teste, G. Zurfluh, "*Algèbre OLAP et langage graphique*", XXIVéme congrès INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID'06), Hammamet (Tunisie), p. 1039-1054, Juin 2006.

[Ravat et al., 2007a] F. Ravat, O. Teste, R. Tournier, G. Zurfluh "*Algebraic and graphic languages for OLAP manipulations*" International Journal of Data Warehousing and Mining, Idea Group, Vol. 4, N° 1, p. 17-46 , January-March 2008 (soumis en Octobre 2006, revisité en Avril 2007, Accepté en Juin 2007).

[Ravat et al., 2007b] F. Ravat, O. Teste, R. Tournier "*OLAP Aggregation Function for Textual Data Warehouse*", 9th International Conference on Enterprise Information Systems (ICEIS'07), Funchal, Madeira - Portugal, 12/06/2007-16/06/2007, Volume DISI, INSTICC Press, p. 151-156, June 2007.

[Ravat et al., 2007c] F. Ravat, O. Teste, R. Tournier, G. Zurfluh "*Integrating Complex Data into a Data Warehouse*" International Conference on Software Engineering and Knowledge Engineering (SEKE'07), Boston, World Scientific Publishing, July 09-11, 2007.

[Ravat et al., 2007d]F. Ravat, O. Teste, R. Tournier, G. Zurfluh "*Querying Multidimensional Databases*", 11th East-European Conference on Advances in Databases and Information Systems (ADBIS'07), Varna (Bulgarie), Springer, LNCS 4690, p. 298-313, September 29 - October 03, 2007.

[Ravat et al., 2007e] F. Ravat, O. Teste, R. Tournier, Gilles Zurfluh "*A Conceptual Model for Multidimensional Analysis of Documents*", 26th International Conference on Conceptual Modeling (ER 2007), Auckland, New Zealand, November 05-11, 2007.

[Ravat et al., 2007f] F. Ravat, O. Teste, G. Zurfluh, "P*ersonnalisation de bases de données multidimensionnelles*", Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'07), Perros-Guirec, p. 121-136, 22-25 Mai 2007.

[Ravat et al., 2007g] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, "*Modèle conceptuel pour l'analyse multidimensionnelle de documents*", Journées Entrepôts de Données et Analyse en ligne (EDA'07), Revue des Nouvelles Technologies de l'Information, Cépaduès Editions, Numéro spécial V. RNTI-B-3, p. 161-175, Juin 2007.

[Ravat et al., 2007h] F. Ravat, O. Teste, R. Tournier, "Analyse multidimensionnelle de documents via des dimensions OLAP". Document numérique, Hermès, Numéro spécial Entreposage de documents et de données semi-structurées, Vol 9, 2007 (à paraître)

[Ravat, 1996] F. Ravat "*OD3 : contribution méthodologique à la conception de bases de données orientées objet réparties*", Thèse de doctorat de l'Université Paul Sabatier, Septembre 2006.

[Riahi, 1998] F. Riahi. "*Elaboration Automatique d'une Base de Données à partir d'Informations Semi-Structurées issues du Web*" 16ème Conférence INFormatique des ORGanisations, des Systèmes d'Information et de Décision (INFORSID'98), Toulouse, INFORSID, Mai 1998.

[Riahi, 2000] F. Riahi. "*Mécanismes pour l'élaboration automatique d'un entrepôt d'informations à partir de documents semi-structurés issus du Web*" Thèse de doctorat de l'Université Paul Sabatier (Toulouse III), Mai 2000.

[Rieu, 1999] D. Rieu, "*Ingénierie des systèmes d'information: bases de données, bases de connaissances et méthodes de conception*" Habilitation à diriger des recherches en Informatique, Institut National Polytechnique de Grenoble, Grenoble, 1999.

[Rijsbergen, 1979] C. Van Rijsbergen "*Information retrieval*", Second Edition, Butterworths, Londres, 1979.

[Rizzi et al., 2006] S. Rizzi, A. Abelló, J. Lechtenbörger, J. Trujillo, "*Research in data warehouse modeling and design: dead or alive?*", 9th ACM Intl. Workshop on Data Warehousing and OLAP (DOLAP'06), Arlington (Virginia, USA), p. 3–10, November 10, 2006.

[Romero & Abelló] O. Romero, A. Abelló, "On the need of a reference algebra for OLAP", 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg (Germany), Springer-Verlag, LNCS 4654, p.99-110, 03-07 Septembre 2007.

[Roques, 2003] P. Roques, "*UML in Practice: The Art of Modeling Software Systems Demonstrated through Worked Examples and Solutions*", John Wiley & Sons, Inc., New York, ISBN: 978-0-470-84831-9, December 2003

[Roques & Vallee, 2004] P. Roques, F. Vallée, "*UML 2 en action : de l'analyse des besoins à la conception J2EE*", Eyrolles (Eds), Collection Architecte logiciel, 3ème édition, ISBN10 : 2-212-11462-1, Juin 2004

[Rouhaud, 2005] O. Rouhaud, "*Interface d'interrogation incrémentale de données multidimensionnelles*" Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

S

[Sallami, 2004] M. Sallami, "*Développement d'une politique d'accès aux bases en constellation*", Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2004.

[Salton & McGill, 1983] G. Salton, M.J. McGill "*Introduction to modern Information Retrieval*" Mc Graw Hill International Book Compagny, 2° edition, September 1983.

[Salton et al., 1997] G. Salton, A. Singhal, M. Mitra, C. Buckley "*Automatic Text Structuring and Summarization*", Information Processing and Management, Vol. 33(2) p.193-207, Janvier 1997.

[Salton, 1971] G. Salton "*The SMART Retrieval System: Experiment in Automatic Document Processing*", Prentice Hall Inc., Englewood Cliffs, 1971.

[Salton, 1989] G. Salton "*Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*", Addison Wesley Publishing Company, 1989.

[Samtani et al., 1998] S. Samtani, M. K. Mohania, V. Kumar, Y. Kambayashi, "*Recent Advances and Research Problems in Data Warehousing*" Advances in Database Technologies, ER '98 Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support and Spatio-Temporal Data Management, Singapore, Lecture Notes in Computer Science 1552, p. 81-92, November 19-20, 1998,

[Sapia et al., 1998] C. Sapia, M. Blaschka, G. Höfling, B. Dinter, "*Extending the E/R Model for the Multidimensional Paradigm*", Advances in Database Technologies, ER '98 Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support and Spatio-Temporal Data Management (ER Workshops), LNCS 1552, Springer, p. 105–116, 1998.

[Schiefer at al., 2002] J. Schiefer, B. List, R. M. Bruckner, "A Holistic Approach for Managing Requirements of Data Warehouse Systems" 8th Americas Conference on Information Systems (AMCIS'02), Dallas (Texas, USA), p. 77-87, August 2002.

[Schneider, 2003] M. Schneider "*Well-formed data warehouse structures*", 5th International Workshop on Design and Management of Data Warehouses (DMDW'03) Berlin (Germany), 8 Septembre 2003.

[Schneider, 2007] M. Schneider "*A general model for the design of data warehouses*" International Journal of Production Economics, In Press, Corrected Proof, (doi:10.1016/j.ijpe.2006.11.027) Available online 19 April 2007.

[Sédes, 1998] F. Sédes "*Bases documentaires – hyperbases. Proposition d'un modèle générique et contribution à la spécification d'un langage pour l'intégration de la manipulation d'informations semi-structurées*" Mémoire d'habilitation à diriger des Recherches, Université Paul Sabatier (Toulouse III), Décembre 1998.

[Sen & Sinha, 2005] A. Sen, A. P. Sinha, "A comparison of data warehousing methodologies" Communication of ACM, Vol. 48(3), p. 79-84, Mars 2005.

[Sheth & Larson, 1990] A. P. Sheth, J. A. Larson "*Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases*" ACM Computer Survey Vol. 22(3), p. 183-236 Septembre 1990.

[Shim et al., 2002] JP. Shim, M. Warkentin, JF Courtney, DJ. Power, R. Sharda, C. Carlson, "*Past, present and future of decision support technology*", Decision Support Systems 33 (2), p. 111-126, June 2002.

[Shukla et al., 1998] A. Shukla, P.M. Deshpande, J.F. Naughton, K. Ramasamy *"Storage Estimation for Multidimensional Aggregates in the Presence of Hierarchies*", 22th International Conference on Very Large Data Bases, (VLDB'96), Mumbai (Bombay, India), p. 522-531, September 3-6, 1996.

[Soulé-Dupuy, 1990] C. Soulé-Dupuy "*Systèmes de recherche d'information : Mécanismes d'indexation et d'interrogation*" Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), 1990.

[Soulé-Dupuy, 2001] C. Soulé-Dupuy "*Bases d'informations textuelles : des modèles aux applications*", mémoire d'habilitation à diriger des recherches, Université Paul Sabatier (Toulouse III), Décembre 2001.

[Soussi et al., 2005] A. Soussi, J. Feki, F. Gargouri, "*Approche semi-automatisée de conception de schémas multidimensionnels valides*", Journées Entrepôts de Données et l'Analyse en ligne (EDA'2005), Revue des Nouvelles Technologies de l'Information, RNTI-B-1, , Cépadues (ed.), p. 71-90, Juin 2005.

[Srivastava & Chen, 1999] J. Srivastava, P.-Y. Chen, "Warehouse Creation - A Potential Roadblock to Data Warehousing". IEEE Transactions on Knowledge and Data Engineering Vol.11(1), p. 118-126, January/February 1999.

[Stern, 1997] Y. Stern "*Les quatre dimensions du document*", Document numérique, Vol. 1(1), Editions Hermes, 1997.

[Sullivan, 2001] D. Sullivan "*Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*", John Wiley & Sons, ISBN 0471399590, February 2001.

T

[Tahi, 2005] A. Tahi, "Bases de données décisionnelles : Fusion de tables multidimensionnelles" Mémoire du Master 2 Recherche 2IH (Informatique, Image et Hypermédia) de l'Université Paul Sabatier (Toulouse III), Juin 2005.

[Teste, 2000] O. Teste " *Modélisation et manipulation d'entrepôts de données historisées*", Thèse de doctorat de l'Université Paul Sabatier, Décembre 2000.

[Theodoratos & Bouzeghoub, 1999] D. Theodoratos, M. Bouzeghoub, "Data Currency Quality Factors in Data Warehouse Design", 1st International Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg (Germany), Paper 15, June 14-15, 1999.

[Theodoratos & Sellis 1997] D. Theodoratos, T. Sellis "*Data warehouse configuration*", 23rd International Conference on Very Large Data Bases (VLDB'97), Athens (Greece), p. 126-135, August 25-29 1997.

[Theodoratos & Sellis 1999] D. Theodoratos, T. Sellis "*Designing Data Warehouses*", Data & Knowledge Engineering, Volume 31 (3) p. 279-301, 1999.

[Tininini, 2003] L. Tininini, "*Querying multidimensional Data*", Chapter IX of "Multidimensional databases: problems and solutions" de M. Rafanelli, Idea Group Publishing, ISBN 1-59140-053-8, pp 252-281, 2003.

[Torlone, 2003] R. Torlone, "*Conceptual Multidimensional Models*", Chapitre III, Multidimensional Databases: Problems and Solutions, Maurizio Rafanelli (Ed.), Idea Publishing Group (IGP), ISBN 1-59140-053-8, p. 69–90, 2003.

[Tournier, 2003] R. Tournier, "Vers un langage de manipulation graphique des bases multidimensionnelles" Mémoire de DEA 2IL (Informatique de l'Image et du Langage) de l'Université Paul Sabatier (Toulouse III), Juin 2003.

[Tournier, 2007] R. Tournier, "Analyse en ligne (OLAP) de documents", Thèse de doctorat en Informatique de l'Université Paul Sabatier (Toulouse III), Décembre 2007 (à paraître).

[Trujillo at al., 2003] S. Luján-Mora, J. Trujillo, I. Song "*Applying UML for designing multidimensional databases and OLAP Applications*", Advanced Topics in Database Research, Vol. 2. Idea Group, p. 13-36, 2003.

[Tryfona et al., 1999] N. Tryfona, F. Busborg, J. G. Borch Christiansen, "*starER: A Conceptual Model for Data Warehouse Design*", 2nd ACM International Workshop on Data Warehousing and OLAP (DOLAP'99), Kansas City (Missouri, USA), p. 3–8, November 6, 1999.

[Tseng & Chou, 2006] F. S.C. Tseng, A. Y.H. Chou, "*The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence*", journal of Decision Support Systems (DSS), vol. 42(2), Elsevier, p. 727–744, Novembre 2006.

[Tsois et al., 2001] A. Tsois, N. Karayannidis, T. K. Sellis, "*MAC: Conceptual data modeling for OLAP*", 3rd International Workshop on Design and Management of Data Warehouses, (DMDW'01), Interlaken (Switzerland), Paper 5, June 4, 2001.

[Tuffery, 1984] M. Tuffery "*Système documentaire, base de données textuelles : le projet ETOILE*", Thèse de doctorat de 3°cycle en Informatique, N°2990, Université Paul Sabatier (Toulouse III), Juin 1984.

U-V

[Vaisman & Mendelzon, 2001] A. A. Vaisman, A. O. Mendelzon, "A temporal query language for OLAP: Implementation and a case study" 8th Biennial Workshop on Data Bases and Programming Languages (DBPL'01), Rome (Italy), Lecture Notes in Computer Science 2397, p. 78-96, 8-10 September 2001.

[Vaisman et al., 2002] A.A. Vaisman, A.O. Mendelzon, W. Ruaro, S.G. Cymerman, "Supporting dimension updates in an OLAP Server", 14th International Conference on Advanced Information Systems Engineering (CaiSE'02), Toronto (Canada), LNCS 2348, p. 67-82, 27-31 May 2002.

[Vassiliadis, 1998] P. Vassiliadis, "*Modeling Multidimensional Databases, Cubes and Cube Operations*". 10th International Conference on Scientific and Statistical Database Management (SSDBM'98), Capri (Italy), p. 53-62, July 1-3, 1998.

[Vassiliadis & Sellis, 1999] P. Vassiliadis, T.K. Sellis, "*A Survey of Logical Models for OLAP Databases*". *SIGMOD Record* 28(4), pp. 64-69, December 1999.

[Vassiliadis et al., 1999] P. Vassiliadis, M. Bouzeghoub, C. Quix, "Towards Quality-Oriented Data Warehouse Usage and Evolution" 11th International Conference on Advanced Information Systems Engineering (CaiSE'99), Heidelberg (Germany), Lecture Notes in Computer Science 1626, Springer, p. 164-179, June 14-18, 1999.

[Vassiliadis et al., 2002] P. Vassiliadis, A. Simitsis, S. Skiadopoulos, "Conceptual modeling for ETL processes" 5th International Workshop on Data Warehousing and OLAP, (DOLAP'02) McLean (VA, USA), p. 14-21, November 8, 2002.

[Vetterli et al., 2000] T. Vetterli, A. Vaduva, M. Staudt, "Metadata Standards for Data Warehousing: Open Information Model vs. Common Warehouse Metamodel" SIGMOD Record Vol. 29(3), p. 68-75, September 2000.

W

[W3C 2000] W3C, "*eXtensible Markup Language (XML) 1.0 (Second Edition)*", W3C Recommendation 6 October 2000 (http://www.w3.org/TR/2000/REC-xml-20001006/), October 2000.

[W3C, 1999] W3C, "HTML 4.01 Specifications", W3C Recommendation 24 December 1999 (http://www.w3.org/TR/1999/REC-html401-19991224/) December 1999.

[Wang et al., 2003] H. Wang, J. Li, Z. He, H. Gao, "*Xaggregation: Flexible Aggregation of XML Data*", 4th International Conference on Advances in Web-Age Information Management (WAIM'03), LNCS 2762, Springer, p. 104–115, Chengdu (China), August 17-19, 2003.

[Wang et al., 2005] Hongzhi Wang, Jianzhong Li, Zhenying He, Hong Gao, "*OLAP for XML Data*", 5th International Conference on Computer and Information Technology (CIT'05), Shanghai (China), p. 233–237, September 21-23, 2005.

[Widom & Ceri, 1996] J. Widom, S.Ceri "*Active Database Systems: Triggers and Rules for Advanced Database Processing*" Morgan Kaufmann, ISBN-10: 1558603042, 1996.

[Widom, 1995] J Widom., "*Research problems in data warehousing*", 4th International Conference on Information and Knowledge Management (CIKM'95), Baltimore - USA, pp 25-30, November 29 – 2 December 2, 1995.

[Widom, 1996] J. Widom, "*The Starburst Active Database Rule System*", IEEE Transaction on Knowledge Data Engineering Vol. 8(4), p. 583-595, Août 1996.

[Wolfe, 2002] J. Wolfe, "*Annotation technologies: A software and research review*", Computers and Composition, Vol. 19(4), p. 471-497, December 2002

[Wrembel & Morzy, 2005] R. Wrembel, T. Morzy, "*Multiversion Data Warehouses: Challenges and Solutions*" IEEE Conference on Computational Cybernetics (ICCC'05), Mauritius, 2005.

[Wu et al., 2004] K. Wu, E. J. Otoo, A. Shoshani, "*On the performance of bitmap indices for high cardinality attributes*" 30th International Conference on Very Large Data Bases (VLDB'04), Toronto (Canada), p. 24-35, 31 Août - 3 Septembre 2004.

X-Y

[Yang & Widom, 1998] J. Yang, J. Widom "*Maintaining Temporal Views Over Non-Temporal Information Sources For Data Warehousing*", 6th International Conference on Extending Database Technology (EDBT'98), Valencia (Spain), p. 389-403, March 23-27 1998.

[Yang & Widom, 2000] J. Yang, J. Widom "*Temporal View Self-Maintenance in a Warehousing Environment*", 7th International Conference on Extending Database Technology (EDBT'00), Konstanz (Germany), p. 395-412, March 27-31 2000.

[Yang et al., 1997] J. Yang , K. Karlapalem, Q. Li "*Algorithms for materialized view design in data warehousing environment*", 23rd International Conference on Very Large Data Bases (VLDB'97), Athens (Greece), p. 136-145, August 25-29 1997.

Z

[Zhou et al., 1996] G. Zhou, R. Hull, R. King "*Generating Data Integration Mediators that Use Materialization*", Journal of Intelligent Information Systems, Vol. 6(2-3), ISSN:0925-9902, p. 199-221, June 1996.

[Zhou et al., 2000] S. Zhou, A. Zhou, X. Tao, Y. Hu "*Hierarchically distributed data warehouse*" 4th International Conference on High Performance Computing in the Asia-Pacific Region, Vol. 2, Issue p.848 – 853, 14-17 Mai 2000.

[Zhuge & Garcia-Molina, 1998] Y. Zhuge, H. Garcia-Molina "*Graph Structured Views and Their Incremental Maintenance*", 14th International Conference on Data Engineering (ICDE'98), Orlando (Florida, USA), p. 116-125, February 23-27 1998.

[Zhuge et al., 1995] Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom "*View Maintenance in a Warehousing Environment*", 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD'95), San Jose (California, USA), p. 316-327, May 22-25 1995.

[Zhuge et al., 1996] Y. Zhuge, H. Garcia-Molina, J. L. Wiener "*The Strobe Algorithms for Multi-Source Warehouse Consistency*", 4th International Conference on Parallel and Distributed Information Systems (PDIS'96), Miami Beach (Florida, USA), p. 146-157, December 18-20, 1996.

[Zhuge et al., 1997] Y. Zhuge, H. Garcia-Molina, J. L. Wiener "*Multiple View Consistency for Data Warehousing*", 13th International Conference on Data Engineering (ICDE'97), Birmingham (U.K.), p. 289-300, April 7-11 1997.

[Zhuge et al., 1998] Y. Zhuge, H. Garcia-Molina, J. L. Wiener "*Consistency Algorithms for Multi-Source Warehouse View Maintenance*", Journal of Distributed and Parallel Databases, Vol. 6(1), p. 7-40, January 1998.

[Zubcoff & Trujillo, 2006] J. J. Zubcoff, J. Trujillo, "*Conceptual Modeling for Classification Mining in Data Warehouses*", 8th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'06), Krakow (Poland), Springer Verlag, LNCS 4081, p. 566-575, September 2006.

SIGLES

SIGLES

BD : Base de Données

BDM : Base de Données Multidimensionnelles

DTD : Document Type Definition ou Définition de Type de Document

ED : Entrepôt de Données

EDO : Entrepôt de DOcument

ETC : Extraction Transformation Chargement

ETL : Extraction Transformation Loading

H-OLAP : Hybrid - On Line Analytical Processing

MD : Magasin de Données

M-OLAP : Multidimensional – On Line Analytical Processing

ODMG : Object Data Management Group

OLAP : On-Line Analytical Processing

O-OLAP : Object– On Line Analytical Processing

R-OLAP : Relational – On Line Analytical Processing

SAD : Système d'Aide à la Décision

SGBD : Système de Gestion de Bases de Données

SI : Système d'Information

SIAD : Système d'Information d'Aide à la Décision

SO : Système Opérant

SP : Système de Pilotage

SRI : Système de Recherche d'Information

TM : Table Multidimensionnelle

Habilitation à Diriger des Recherches

Modèles et outils pour la conception et la manipulation de systèmes d'aide à la décision

Franck RAVAT

IRIT-UT1 Equipe SIG-ED

Résumé

Nos travaux se situent dans le cadre des Systèmes d'Aide à la Décision (SAD). Au début de nos travaux, nous étions en présence de solutions d'ordre technique pour l'alimentation des SAD (vues matérialisées) ainsi que de quelques solutions parcellaires pour la modélisation et les manipulations multidimensionnelles. Durant ces dernières années, notre objectif a été d'offrir une solution globale pour la conception et la manipulation de SAD. Dans un premier temps, nous avons identifié deux espaces de stockage pour les données décisionnelles : un Entrepôt de Données (ED) et des Magasins de Données (MD). Un ED centralise et historise les données issues des sources de production et chaque MD présente les données à un décideur pour faciliter ses prises de décisions.

Pour les ED, notre objectif a été d'apporter des solutions pour la modélisation de l'évolution des données décisionnelles (extension de modèle objet) et pour l'intégration de données textuelles sans en fixer le schéma à priori. Pour les MD, nous avons proposé un modèle multidimensionnel de base avec différentes extensions répondant aux besoins des décideurs. Ces extensions permettent de prendre en compte la gestion d'indicateurs et de données textuelles, l'évolution temporelle (versions), la cohérence des données et de ses analyses (contraintes sémantiques), l'intégration et la capitalisation de l'expertise des décideurs (annotations) ainsi que la personnalisation des schémas multidimensionnels (poids). Ces travaux ont été complétés par la proposition d'une démarche de conception qui présente l'avantage de prendre en compte les besoins des décideurs et les sources de données. Cette démarche permet de modéliser aussi bien l'aspect statique (données décisionnelles) que l'aspect dynamique (processus d'alimentation du SAD).

D'un point de vue manipulation des données, nous avons proposé une algèbre complétée d'un langage graphique orienté décideur et d'un langage déclaratif. Nos propositions ont été validées par la participation à différents projets ainsi que le co-encadrement de 5 thèses de doctorat et le suivi de travaux de plusieurs Master Recherche.

Mots clés

Entrepôt de données, Magasin de données, Modèles et langages OLAP, Méthode de conception